

Phase-space negativity as a computational resource for quantum kernel methods

Ulysse Chabaud¹, Roohollah Ghobadi², Salman Beigi³, and Saleh Rahimi-Keshari⁴

¹DIENS, École Normale Supérieure, PSL University, CNRS, INRIA, 45 rue d'Ulm, Paris 75005, France

²Institute for Quantum Science and Technology, University of Calgary, Calgary, AB, T2N 1N4, Canada

³School of Mathematics, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746, Tehran, Iran

⁴School of Physics, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5531, Tehran, Iran

Quantum kernel methods are a proposal for achieving quantum computational advantage in machine learning. They are based on a hybrid classical-quantum computation where a function called the quantum kernel is estimated by a quantum device while the rest of computation is performed classically. Quantum advantages may be achieved through this method only if the quantum kernel function cannot be estimated efficiently on a classical computer. In this paper, we provide sufficient conditions for the efficient classical estimation of quantum kernel functions for bosonic systems. These conditions are based on phase-space properties of data-encoding quantum states associated with the quantum kernels: negative volume, non-classical depth, and excess range, which are shown to be three signatures of phase-space negativity. We consider quantum optical examples involving linear-optical networks with and without adaptive non-Gaussian measurements, and investigate the effects of loss on the efficiency of the classical simulation. Our results underpin the role of the negativity in phase-space quasi-probability distributions as an essential resource in quantum machine learning based on kernel methods.

ulysse.chabaud@inria.fr
farid.ghobadi80@gmail.com
salman.beigi@gmail.com
s.rahimik@gmail.com

1 Introduction

Small-scale quantum devices with a few hundred qubits [1, 2] represent a novel paradigm for applications in quantum simulation [3], quantum chemistry [4], and quantum machine learning [5, 6]. Despite their relatively small scale, there are strong evidences based on complexity-theoretic arguments, that simulating the sampling statistics of these devices is beyond the reach of classical computers [7–10]. However, determining why quantum devices can outperform classical ones remains a fundamental question in quantum information science. In particular, identifying the necessary quantum resources and understanding the effects of errors on them is a complex challenge.

One approach to address this challenge, especially for quantum sampling problems, has been based on the use of quasi-probability distributions [11–17]. In this approach, a classical description of a quantum experiment is possible as long as one can find probabilistic descriptions for the input state, the evolution, and the measurement in terms of non-negative quasi-probability distributions. Also, the relation between the negative volume of quasi-probability distributions and the computational overhead in estimating an output probability of a quantum circuit using those distributions has been investigated [16, 17].

In this paper, we build upon these results to quantify the sampling complexity in estimating *quantum kernels* in terms of the negativity of the associated quasi-probability distributions. Quantum kernel methods have been proposed as a promising near-term application of quantum computers [18], and the non-classicality of quantum kernels has been previously investigated in

[19], based on phase-space inequalities from [20]. Hereafter, we go beyond these conceptual characterizations and present sufficient conditions for efficient classical estimation of kernel functions in continuous-variable quantum machine learning. Informally, we show that one can achieve the same performance as a quantum circuit by using classical Monte Carlo-type sampling techniques if the negativity of the phase-space quasiprobability distributions (PQD) of data-encoding quantum states is limited. This condition is independent of the measurement strategy and holds even if sampling from the output probability distribution of the circuit is classically hard, such as in boson sampling problems.

In light of this, we identify the negativity of PQDs as a necessary resource to achieve quantum computational speedups in machine learning based on kernel methods. More precisely, we obtain classical algorithms with fine-grained sample complexity based on specific phase-space properties, namely negative volume [21, 22], non-classical depth [23, 24] and excess ranges. While the negative volume is a direct measure of the negativity of a PQD, the non-classical depth indicates the amount of thermal noise necessary to render a PQD non-negative, and we further show that excess range of PQDs can also be understood a signature of phase-space negativity. As illustrating examples, we discuss efficient classical estimation of quantum kernel functions that are based on Gaussian states, non-Gaussian output states of linear optical networks, and partially measured Gaussian states, including adaptive measurement strategies.

Our results involve a new approach to classical estimation of quantum kernel functions beyond probability estimation [16, 17] that takes into account the structure of state preparation. A surprising consequence is that estimating quantum kernel functions based on a large class of partially measured Gaussian states can be done efficiently by classical computers despite these states being highly non-Gaussian (see Theorem 1).

The structure of the paper is as follows: in Section 2, we set the stage and provide some background on quantum kernel methods; in Section 3, we derive general expressions for quantum kernel functions for bosonic systems based on PQDs (see Lemma 1); in Section 4, we introduce classical Monte Carlo sampling methods and derive

two phase-space-inspired classical algorithms for estimating quantum kernel functions (see Algorithms 1 and 2), together with rigorous performance guarantees; we illustrate the flexibility and applicability of our algorithms in Section 5, by applying them to several examples of high significance in bosonic quantum information processing, identifying the relevant quantum computational resources in each setting; we conclude in Section 6.

2 Quantum Kernel Methods

In quantum machine learning, we often model an unknown function f of some classical data x by $f(x) = \text{Tr}[\rho(x)O]$ where $\rho(x)$ is a quantum state depending on the input x , and O is a quantum observable.

A general encoding of classical data x to a quantum state $\rho(x)$ over m subsystems based on a quantum map may take the form

$$x \mapsto \rho(x) = \frac{\text{Tr}_k[(\Pi(x) \otimes \mathbb{I}_m)U(x)\rho_{\text{in}}(x)U(x)^\dagger]}{\text{Tr}[(\Pi(x) \otimes \mathbb{I}_m)U(x)\rho_{\text{in}}(x)U(x)^\dagger]}, \quad (1)$$

where the classical data x may be encoded in a density operator $\rho_{\text{in}}(x)$ describing an initial quantum state, a unitary operator $U(x)$ describing a quantum circuit, and a positive operator-valued measure (POVM) element $\Pi(x)$ describing the measurement, and where the subscript k denotes that the first k subsystems are being measured. The denominator is for normalisation of the post-measurement state

$$\rho_{\Pi(x)} := \text{Tr}_k[(\Pi(x) \otimes \mathbb{I}_m)U(x)\rho_{\text{in}}(x)U(x)^\dagger]. \quad (2)$$

It is often the case that classical data is encoded in the input state or the circuit parameters only, since these are the parameters that can be easily varied in practice [25]. Note also that the dependency of ρ_{in} or Π on the classical data x can be moved to $U(x)$ without loss of generality. Similarly, the initial state ρ_{in} can be chosen as a tensor product state without loss of generality by changing U . That being said, we allow for very general encoding strategies and let the input, evolution and measurement potentially depend on classical data x .

In this setting, the encoding map $x \mapsto [\rho_{\text{in}}(x), U(x), \Pi(x)]$ is known, while the observable O is unknown. The goal is to approximate

$f(x)$ given a training dataset $\mathcal{D} = \{(x_i, y_i = f(x_i))\}_{i=1}^n$. To this end, we look for an observable $\tilde{O}$ that minimizes the cost function

$$\frac{1}{n} \sum_{i=1}^n c(\text{Tr}(\rho(x_i)\tilde{O}), y_i) + \lambda \text{Tr}(\tilde{O}^2). \quad (3)$$

Here, $c(\cdot, \cdot) \geq 0$ is a cost function which measures the distance of the predicted value $\text{Tr}(\rho(x_i)\tilde{O})$ and the true value y_i . The second term with the regularization parameter $\lambda > 0$ ensures avoiding overfitting. This regularization term penalizes complex hypotheses that best match the training dataset but do not provide a good prediction for arbitrary inputs.

Invoking the representer theorem, it can be shown that the optimal observable $\tilde{O}$ that minimizes (3) can be written as [26, 27]

$$\tilde{O} = \sum_{i=1}^n \alpha_i \rho(x_i), \quad (4)$$

where α_i 's are real numbers. In this case, the optimal approximating function equals

$$\tilde{f}(x) = \sum_{i=1}^n \alpha_i K(x, x_i), \quad (5)$$

where we have introduced the *kernel function*

$$K(x, x') = \text{Tr}[\rho(x)\rho(x')] = \frac{\text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]}{\text{Tr}[\rho_{\Pi(x)}]\text{Tr}[\rho_{\Pi(x')}]}, \quad (6)$$

where the right hand side is obtained using Eqs. (1) and (2). This kernel function can be viewed as a measure of similarity between data points.

According to Eq. (5), in order to find the optimal function $\tilde{f}(x)$ we only need to compute the values of the kernel function; given two data points x, x' , we need to be able to compute the overlap of $\rho(x), \rho(x')$. The quantum kernel methods [28, 29] are hybrid classical-quantum machine learning techniques which involve estimating these overlaps quantumly to within an inverse-polynomial additive error in the size of the corresponding quantum states, while the rest of the computation, i.e., computing the coefficients α_i 's from the kernel values $K(x, x_i)$ is done classically. In particular, the quantum part of the computation consists in generating a polynomial number of copies of the states $\rho(x), \rho(x')$ and using these copies to estimate their overlap

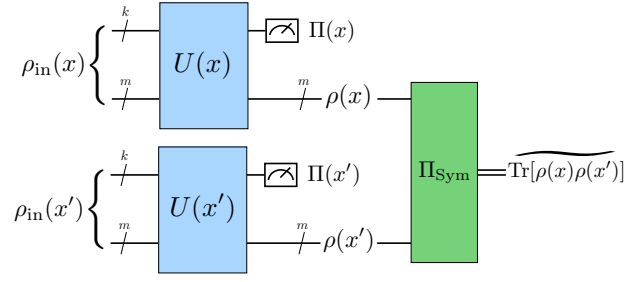


Figure 1: Quantum estimation of quantum state overlap. The green block Π_{sym} represents the projection of $\rho(x) \otimes \rho(x')$ onto the symmetric subspace. Repeating that projection N times allows to estimate the overlap $K(x, x')$ up to inverse-polynomial precision (in N) with exponentially small failure probability, by computing the frequency of successful projection.

up to inverse-polynomial precision, for instance using the SWAP test [30], which effectively implements a projection Π_{sym} onto the symmetric subspace (see Fig. 1). In the case of an encoding involving quantum measurements, this is only efficient when the probability to generate the states $\rho(x), \rho(x')$ is at least inverse-polynomially large, i.e., $\text{Tr}[\rho_{\Pi(x)}] \geq 1/\text{poly}(m)$, $\text{Tr}[\rho_{\Pi(x')}] \geq 1/\text{poly}(m)$. When this is the case we say that the encoding is *quantum-efficient*. As such, quantum computational advantage using quantum kernel methods may only come from the estimation of kernel values in the case of quantum-efficient encodings.

In the rest of this paper, we show using phase-space techniques that for a large family of quantum-efficient encoding schemes $x \mapsto \rho(x)$, the kernel values (6) can be estimated efficiently classically up to the same precision as can be done quantumly. When that is the case, there can be no computational advantage from quantum kernel methods.

3 Kernel functions in phase space

Let ρ be a quantum state of an m -mode bosonic system. Then ρ can be represented as [11, 31]:

$$\rho = \pi^m \int_{\mathbb{C}^m} d^{2m}\alpha W_\rho^{(s)}(\alpha) \Delta^{(-s)}(\alpha), \quad (7)$$

where α is a row vector of m complex numbers, where $d^{2m}\alpha = d^m \Re(\alpha) d^m \Im(\alpha)$, and where the

$\mathbf{s}$ -ordered phase-space quasi-probability distributions [($\mathbf{s}$)-PQDs] are defined by

$$W_{\rho}^{(\mathbf{s})}(\boldsymbol{\alpha}) = \text{Tr}[\rho \Delta^{(\mathbf{s})}(\boldsymbol{\alpha})], \quad (8)$$

with phase-space point operators

$$\Delta^{(\mathbf{s})}(\boldsymbol{\alpha}) = \int_{\mathbb{C}^m} \frac{d^{2m}\boldsymbol{\xi}}{\pi^{2m}} D(\boldsymbol{\xi}) e^{\boldsymbol{\xi} \mathbf{s} \boldsymbol{\xi}^\dagger / 2} e^{\boldsymbol{\alpha} \boldsymbol{\xi}^\dagger - \boldsymbol{\xi} \boldsymbol{\alpha}^\dagger}. \quad (9)$$

Here, $\boldsymbol{\alpha}, \boldsymbol{\xi}$ are row vectors of m complex numbers, the diagonal matrix $\mathbf{s} = \text{diag}(s_1, s_2, \dots, s_m)$ specifies the ordering for each mode, and $D(\boldsymbol{\xi}) = \exp(\boldsymbol{\xi} \mathbf{a}^\dagger - \mathbf{a} \boldsymbol{\xi}^\dagger)$ is the m -mode displacement operator with $\mathbf{a} = (a_1, \dots, a_m)$ and $\mathbf{a}^\dagger = (a_1^\dagger, \dots, a_m^\dagger)^T$ being vectors of annihilation and creation operators, respectively. Note that these formulas are formally valid for more general ordering matrices, but we restrict to diagonal ones for simplicity.

Using $\int d^{2m}\boldsymbol{\alpha} \exp(\boldsymbol{\alpha} \boldsymbol{\xi}^\dagger - \boldsymbol{\xi} \boldsymbol{\alpha}^\dagger) = \pi^{2m} \delta^{2m}(\boldsymbol{\xi})$ and Eq. (9), one can verify that the ($\mathbf{s}$)-PQD

of density operators are normalized. In general, they can take negative values or do not represent probabilities of mutually exclusive events, which is why they are named quasi-probability distributions.

When $\mathbf{s}$ is the identity matrix $\mathbf{I}$, the ($\mathbf{s}$)-PQD becomes the Glauber–Sudarshan representation which is a highly singular distribution for most quantum states. Only for classical states, which can be viewed as statistical mixtures of coherent states, the Glauber–Sudarshan P function is non-negative. Other special cases are the Wigner function for $\mathbf{s} = \mathbf{0}$ that takes on negative values for some non-classical states, and the Husimi function for $\mathbf{s} = -\mathbf{I}$, in which case $W_{\rho}^{(-\mathbf{I})}(\boldsymbol{\alpha}) = \langle \boldsymbol{\alpha} | \rho | \boldsymbol{\alpha} \rangle / \pi^m$, where $|\boldsymbol{\alpha}\rangle$ denotes an m -mode coherent state. We note that for the Husimi Q function we always have $0 \leq W_{\rho}^{(-\mathbf{I})}(\boldsymbol{\alpha}) \leq 1/\pi^m$.

These phase-space quasi-probability distributions allow us to obtain useful expressions for quantum kernel functions. Denoting by $\oplus$ the direct sum of matrices we have:

Lemma 1. *The kernel function $K(x, x') = \text{Tr}[\rho(x)\rho(x')]$ can be expressed as:*

$$K(x, x') = \pi^m \int_{\mathbb{C}^m} d^{2m}\boldsymbol{\gamma} W_{\rho(x')}^{(-\mathbf{s})}(\boldsymbol{\gamma}) W_{\rho(x)}^{(\mathbf{s})}(\boldsymbol{\gamma}), \quad (10)$$

for all $\mathbf{s} = \text{diag}(s_1, s_2, \dots, s_m)$. For a general encoding $x \mapsto \rho(x) = \rho_{\Pi(x)} / \text{Tr}[\rho_{\Pi(x)}]$, it is given by $K(x, x') = \frac{\text{Tr}[\rho_{\Pi(x)} \rho_{\Pi(x')}]}{\text{Tr}[\rho_{\Pi(x)}] \text{Tr}[\rho_{\Pi(x')}]}$, where

$$\text{Tr}[\rho_{\Pi(x)}] = \pi^k \int_{\boldsymbol{\alpha} \in \mathbb{C}^k} d^{2k}\boldsymbol{\alpha} W_{\Pi(x)}^{(-\mathbf{u})}(\boldsymbol{\alpha}) \int_{\boldsymbol{\gamma} \in \mathbb{C}^m} d^{2m}\boldsymbol{\gamma} W_{U(x)\rho_{\text{in}}(x)U(x)^\dagger}^{(\mathbf{u} \oplus \mathbf{s})}(\boldsymbol{\alpha}, \boldsymbol{\gamma}), \quad (11)$$

$$\text{Tr}[\rho_{\Pi(x')}] = \pi^k \int_{\boldsymbol{\beta} \in \mathbb{C}^k} d^{2k}\boldsymbol{\beta} W_{\Pi(x')}^{(-\mathbf{v})}(\boldsymbol{\beta}) \int_{\boldsymbol{\gamma} \in \mathbb{C}^m} d^{2m}\boldsymbol{\gamma} W_{U(x')\rho_{\text{in}}(x')U(x')^\dagger}^{(\mathbf{v} \oplus \mathbf{t})}(\boldsymbol{\beta}, \boldsymbol{\gamma}), \quad (12)$$

for all $\mathbf{s} = \text{diag}(s_1, s_2, \dots, s_m)$, $\mathbf{u} = \text{diag}(u_1, u_2, \dots, u_k)$, $\mathbf{t} = \text{diag}(t_1, t_2, \dots, t_m)$, and $\mathbf{v} = \text{diag}(v_1, v_2, \dots, v_k)$, and where

$$\begin{aligned} \text{Tr}[\rho_{\Pi(x)} \rho_{\Pi(x')}] &= \pi^{m+2k} \int_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{C}^k} d^{2k}\boldsymbol{\alpha} d^{2k}\boldsymbol{\beta} W_{\Pi(x)}^{(-\mathbf{u})}(\boldsymbol{\alpha}) W_{\Pi(x')}^{(\mathbf{v})}(\boldsymbol{\beta}) \\ &\quad \times \int_{\boldsymbol{\gamma} \in \mathbb{C}^m} d^{2m}\boldsymbol{\gamma} W_{U(x)\rho_{\text{in}}(x)U(x)^\dagger}^{(\mathbf{u} \oplus \mathbf{s})}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) W_{U(x')\rho_{\text{in}}(x')U(x')^\dagger}^{(-\mathbf{v} \oplus -\mathbf{s})}(\boldsymbol{\beta}, \boldsymbol{\gamma}). \end{aligned} \quad (13)$$

In the case of a unitary encoding $x \mapsto \rho(x) = U(x)\rho_{\text{in}}(x)U(x)^\dagger$, this simplifies to:

$$K(x, x') = \pi^m \int_{\mathbb{C}^m} d^{2m}\boldsymbol{\gamma} W_{U(x)\rho_{\text{in}}(x)U(x)^\dagger}^{(\mathbf{s})}(\boldsymbol{\gamma}) W_{U(x')\rho_{\text{in}}(x')U(x')^\dagger}^{(-\mathbf{s})}(\boldsymbol{\gamma}). \quad (14)$$

Proof. The expression in Eq. (10) is a direct

consequence of Eq. (7). For the expressions in

Eqs. (11), (12) and (13), let σ be a density operator over $k + m$ modes, Π a POVM element over k modes and $\rho := \text{Tr}_k[(\Pi \otimes I_m)\sigma]$. For all $\gamma \in \mathbb{C}^m$, all $\mathbf{s} = \text{diag}(s_1, \dots, s_m)$ and all $\mathbf{u} = \text{diag}(u_1, \dots, u_k)$,

$$\begin{aligned} W_\rho^{(\mathbf{s})}(\gamma) &= \text{Tr}[\rho \Delta^{(\mathbf{s})}(\gamma)] \\ &= \text{Tr}[\text{Tr}_k[(\Pi \otimes I_m)\sigma] \Delta^{(\mathbf{s})}(\gamma)] \\ &= \text{Tr}[\sigma(\Pi \otimes \Delta^{(\mathbf{s})}(\gamma))] \\ &= \pi^k \int_{\alpha \in \mathbb{C}^k} d^{2k} \alpha W_\Pi^{(-\mathbf{u})}(\alpha) \\ &\quad \times \text{Tr}[\sigma(\Delta^{(\mathbf{u})}(\alpha) \otimes \Delta^{(\mathbf{s})}(\gamma))] \\ &= \pi^k \int_{\alpha \in \mathbb{C}^k} d^{2k} \alpha W_\Pi^{(-\mathbf{u})}(\alpha) W_\sigma^{(\mathbf{u} \oplus \mathbf{s})}(\alpha, \gamma), \end{aligned} \quad (15)$$

where we used Eq. (7) in the fourth line and Eq. (8) in the first and last lines. With Eq. (2), integrating this relation for $W_{\rho_{\Pi(x)}}^{(\mathbf{s})}$ or $W_{\rho_{\Pi(x')}}^{(\mathbf{t})}$ yields Eq. (11) or (12), respectively, while using this relation twice to expand $W_{\rho_{\Pi(x)}}^{(\mathbf{s})}$ and $W_{\rho_{\Pi(x')}}^{(-\mathbf{s})}$ yields Eq. (13). Finally, the last expression in Eq. (14) is a direct consequence of Eq. (10). $\square$

Note that the above Lemma involves phase-space representations of POVM elements that are not necessarily trace-class operators. These should be understood as defined formally in the sense of distributions, i.e., by their inner product with classes of well-behaved functions, which is how they will be employed hereafter.

The rest of the paper is devoted to showing how the general expressions obtained in Lemma 1 enable direct estimation of kernel functions using classical Monte Carlo methods, when various assumptions on the $(\mathbf{s})$ -PQDs are involved.

4 Classical estimation of kernel functions

In this section, we describe two classical algorithms using Monte Carlo sampling methods to estimate kernel functions based on the expressions obtained in Lemma 1. The first Algorithm 1 is obtained by treating kernel estimation as a probability estimation task, and following the approach of [16, 17, 32] (see Fig. 2). The second Algorithm 2 is a non-trivial generalisation which takes advantage of the specifics of state preparation, and uses the first algorithm as a subroutine (see Fig. 3).

4.1 Monte Carlo estimation

Monte Carlo methods are standard algorithms for estimating classical expectation values: given a probability density function $y \mapsto P(y)$ and a bounded function f over $\mathbb{R}^n$, the expectation value $\mathbb{E}_P[f] = \int_{\mathbb{R}^n} f(y) P(y) dy$ over the probability density P can be estimated by computing the mean $\frac{1}{N} \sum_{j=1}^N f(y_j)$ of the function f over a finite number N of samples $y_1, \dots, y_N$ from the probability density P . The error associated with this estimation and the probability of failure are related through Hoeffding's inequality [33]:

$$\Pr \left[\left| \frac{1}{N} \sum_{j=1}^N f(y_j) - \mathbb{E}_P[f] \right| > \epsilon \right] \leq 2 \exp \left(- \frac{2N\epsilon^2}{\mathcal{R}(f)^2} \right), \quad (16)$$

where $\mathcal{R}(f) = \max_{y \in \mathbb{R}^n} f(y) - \min_{y \in \mathbb{R}^n} f(y)$ is the range of the function f . Hence, if the number of samples satisfies $N \geq \frac{1}{2\epsilon^2} \mathcal{R}(f)^2 \ln \left(\frac{2}{\delta} \right)$, then with probability at least $1 - \delta$ the classical estimate of the expectation value $\mathbb{E}_P[f]$ has additive error less than ϵ .

In general, the kernel functions we wish to estimate are of the form $\text{Tr}[\rho A]$, where ρ and A are bounded, positive operators (see Eq. (10)). In this case, the kernel function can be viewed as an expectation value over PQDs rather than true probability distributions. A simple extension of the above method allows us to estimate such quantities: following the approach of [16, 17, 32], given the $(\mathbf{t})$ -PQD, $\mu \mapsto W_\rho^{(\mathbf{t})}(\mu)$ over $\mathbb{C}^n$, representing an operator ρ , we define the probability distribution

$$P(\mu) := \frac{1}{\mathcal{N}(W_\rho^{(\mathbf{t})})} |W_\rho^{(\mathbf{t})}(\mu)|, \quad (17)$$

where $\mathcal{N}(W_\rho^{(\mathbf{t})}) := \int_{\mathbb{C}^n} d^{2n} \mu |W_\rho^{(\mathbf{t})}(\mu)|$ is the *negative volume*, a measure of negativity in the $(\mathbf{t})$ -PQD. We note that $\mathcal{N}(W_\rho^{(\mathbf{t})}) = 1$ if the distribution is non-negative. Next, using the quasi-probability distribution $W_A^{(-\mathbf{t})}(\mu)$ representing operator A , we introduce the estimator

$$E(\mu) := \pi^n \mathcal{N}(W_\rho^{(\mathbf{t})}) \text{sgn}[W_\rho^{(\mathbf{t})}(\mu)] W_A^{(-\mathbf{t})}(\mu), \quad (18)$$

where $\text{sgn}[\cdot] = \pm 1$ is the sign function. By construction, the above quantity provides an unbiased estimator for $\text{Tr}[\rho A]$ through Eq. (7):

$$\mathbb{E}_P[E] = \int_{\mathbb{C}^n} d^{2n} \mu P(\mu) E(\mu) = \text{Tr}[\rho A]. \quad (19)$$

Therefore, assuming that it is possible to evaluate E (and in particular to compute $\mathcal{N}(W_\rho^{(t)})$) and to generate samples from the probability distribution P efficiently classically, $\text{Tr}[\rho A]$ can be estimated using the following procedure: with input (t) -PQDs $W_\rho^{(t)}$ and $W_A^{(-t)}$ for a density operator ρ and a bounded, positive operator A , respectively,

- (i) randomly sample N outcomes $\mu_1, \dots, \mu_N$ from the probability density $P(\mu)$, defined by Eq. (17);
- (ii) using Eq. (18), compute the corresponding values of the estimator $E(\mu_1), \dots, E(\mu_N)$;
- (iii) output the sample mean $\frac{1}{N} \sum_{j=1}^N E(\mu_j)$.

The error associated with this estimation and the probability of failure are once again related through Hoeffding's inequality (16), for the estimator E . Hence, if the number of samples satisfies

$$N \geq \frac{1}{2\epsilon^2} \mathcal{R}(E)^2 \ln\left(\frac{2}{\delta}\right), \quad (20)$$

then with probability at least $1 - \delta$ the classical estimate of the expectation value has additive error less than ϵ . Here, the range of the estimator is bounded as

$$\mathcal{R}(E) \leq 2\mathcal{N}(W_\rho^{(t)})\mathcal{R}(W_A^{(-t)}), \quad (21)$$

with $\mathcal{R}(W_A^{(-t)}) = \pi^n [\max_\mu W_A^{(-t)}(\mu) - \min_\mu W_A^{(-t)}(\mu)]$. This implies that the complexity of the estimation procedure, determined by the number of samples to achieve a desired precision, depends directly on the range of the function $W_A^{(-t)}$ and the negative volume of the PQD $W_\rho^{(t)}$. Note that $\pi^n \max_\mu |W_A^{(-t)}(\mu)| \leq \mathcal{R}(W_A^{(-t)}) \leq 2\pi^n \max_\mu |W_A^{(-t)}(\mu)|$, so we can alternatively use the extremal values rather than the range without affecting the scaling of the sample complexity.

Using this estimation procedure for $\text{Tr}[\rho A]$ to estimate the overlaps in Lemma 1 provides two classical algorithms for kernel estimation.

Algorithm 1. *Choosing $\rho = \rho(x)$ and $A = \rho(x')$, the above estimation procedure provides an additive estimate of the kernel $K(x, x') = \text{Tr}[\rho(x)\rho(x')]$.*

We provide a schematic depiction of this algorithm in Fig. 2. The correctness of Algorithm 1 is a direct consequence of Eq. (10) in Lemma 1 and its efficiency is given by Eqs. (20) and (21), which can be optimized by the choice of (s) -PQD.

Alternatively, we may estimate independently $\text{Tr}[\rho_{\Pi(x)}]$, $\text{Tr}[\rho_{\Pi(x')}]$ and $\text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]$:

Algorithm 2. (i) *Choosing ρ as the k -mode partial trace of $U(x)\rho_{\text{in}}(x)U(x)^\dagger$ over the last m modes and $A = \Pi(x)$, the above estimation procedure provides an additive estimate of $\text{Tr}[\rho_{\Pi(x)}]$.*

(ii) *Choosing ρ as the k -mode partial trace of $U(x')\rho_{\text{in}}(x')U(x')^\dagger$ over the last m modes and $A = \Pi(x')$, the above estimation procedure provides an additive estimate of $\text{Tr}[\rho_{\Pi(x')}]$.*

(iii) *Choosing $\rho = \sigma(x, x')$ as the $(2k)$ -mode overlap of the last m modes of the $(k + m)$ -mode states $U(x)\rho_{\text{in}}(x)U(x)^\dagger$ and $U(x')\rho_{\text{in}}(x')U(x')^\dagger$ and choosing $A = \Pi(x) \otimes \Pi(x')$, the above estimation procedure provides an additive estimate of $\text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]$.*

(iv) *Computing the ratio of these three estimates provides an additive estimate of the kernel $K(x, x') = \text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}] / (\text{Tr}[\rho_{\Pi(x)}]\text{Tr}[\rho_{\Pi(x')}])$.*

We provide a schematic depiction of this algorithm in Fig. 3. The correctness of Algorithm 2 is a consequence Eqs. (11), (12) in Lemma 1 for the estimations of the first two terms and of Eq. (13) in Lemma 1 for the estimation of the third term. Once again, the efficiency of each step is given by Eqs. (20) and (21), and can be optimized by the choice of (s) -PQD.

Note that in Algorithm 1, $\rho(x)$ is used as a source of samples, while the estimator is mainly built from $\rho(x')$ (see Fig. 2). On the other hand, in Algorithm 2, both $U(x)\rho_{\text{in}}(x)U(x)^\dagger$ and $U(x')\rho_{\text{in}}(x')U(x')^\dagger$ are used as sources of samples, while the estimators are mainly built from $\Pi(x)$ and $\Pi(x')$ (see Fig. 3). The first method works well for generic cases, while the second method takes the advantage of the specific state preparation. We derive generic conditions for the efficiency of both methods in the next section, and we give concrete applications of these algorithms in Section 5.

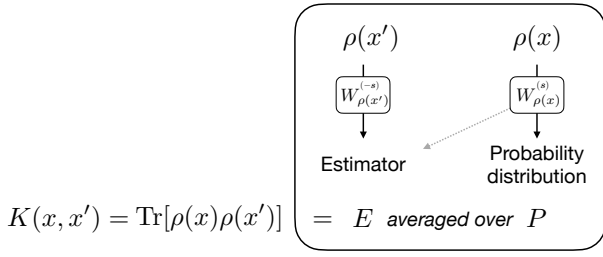


Figure 2: Classical estimation of quantum state overlap, following Algorithm 1. The dashed arrow indicates that the negativity information of $W_{\rho(x')}^{(s)}$ is also being used to define the estimator E according to Eq. (18).

4.2 Conditions for efficiency

The efficiency of the previous methods is based on assuming that all the PQDs and their negative volume can be efficiently computed, and classical efficient sampling from the probability density is possible. However, if efficient computation of the PQDs of the data encoding states is not possible, one may express the PQDs in terms of the initial product states and the transition function describing the encoding circuit, as we show in Appendix A.

We now present sufficient conditions for the efficiency of the above estimation procedures. Following either method, we obtain a classical estimate of the kernel function $K(x, x')$ with additive error $\epsilon = 1/\text{poly}(m)$, with an exponentially small probability of failure, using $N = \text{poly}(m)$ samples, provided that both the negative volume $\mathcal{N}(W_{\rho}^{(t)})$ and the range $\mathcal{R}(W_A^{(-t)})$, bounding the range of the estimator $\mathcal{R}(E)$ as in Eq. (21), are polynomial in m . Thus, to check this condition, one should look for ordering parameters such that $\mathcal{R}(E)$ is minimized.

Notice that having a nonnegative PQD does not necessarily imply efficient classical estimation of the kernel function, since the other contributing range factor must also be considered in the minimization. In fact, the range of PQD is also related to the negativity involved in kernel estimation experimental procedures. As A is a bounded positive operator, it can be viewed as a POVM element of a two-outcome measurement $\{A, \bar{A} = I - A\}$. Therefore, estimation of $\text{Tr}[\rho\bar{A}]$ results in estimation of the complement result of the measurement, $\text{Tr}[\rho A] = 1 - \text{Tr}[\rho\bar{A}]$. How-

ever, since $W_{\bar{A}}^{(-t)}(\mu) = 1/\pi^m - W_A^{(-t)}(\mu)$, we can see that if the range of $W_A^{(-t)}(\mu)$ is greater than $1/\pi^m$, then either it takes negative values, or the PQD $W_{\bar{A}}^{(-t)}(\mu)$ involved in the estimation of $\text{Tr}[\rho\bar{A}]$ becomes negative. In other terms, the excess range of a PQD is a signature of phase-space negativity contributing to the sample complexity of our classical algorithms.

To analyze how the sampling complexity depends on the ordering parameters, we derive an upper bound on the range $\mathcal{R}(W_A^{(-t)})$. Let us first consider a single-mode operator A_1 with PQD $W_{A_1}^{(-t)}(\mu)$. The spectral decomposition of the single-mode Hermitian operator $\Delta^{(-t)}(\mu)$ is given by [11]

$$\Delta^{(-t)}(\mu) = \frac{2}{\pi(1+t)} \sum_{n=0}^{\infty} \left(\frac{t-1}{t+1} \right)^n |n, \mu\rangle \langle n, \mu|, \quad (22)$$

where $|n, \mu\rangle = D(\mu)|n\rangle$ are displaced number states. If $t \geq 0$, we have $(t-1)/(t+1) \leq 1$. As the operator norm $\|A_1\| \leq 1$, the maximum and the minimum values of the single-mode $(-t)$ -PQD are determined by the largest and the smallest eigenvalues in Eq. (22):

$$-\frac{2(1-t)}{\pi(1+t)^2} \leq W_{A_1}^{(-t)}(\mu) \leq \frac{2}{\pi(1+t)}. \quad (23)$$

Therefore, for $t = 0$ the values of the Wigner function are between $-2/\pi$ and $2/\pi$, and as t decreases the length of the interval becomes smaller, down to $1/\pi$ for the Husimi function when $t = 1$. Notice that for $t < 0$ the operator $\Delta^{(-t)}(\mu)$ has infinite eigenvalues, and therefore the values of $(-t)$ -PQD are not bounded in general.

Generalizing Eq. (23) to the m -mode case, we have $\Delta^{(-t)}(\mu) = \otimes_{j=1}^m \Delta^{(-t_j)}(\mu_j)$, and for $t_j \geq 0$,

$$\left(\frac{t_{\min} - 1}{t_{\min} + 1} \right) \prod_{j=1}^m \frac{2}{\pi(1+t_j)} \leq W_A^{(-t)}(\gamma) \leq \prod_{j=1}^m \frac{2}{\pi(1+t_j)},$$

where $t_{\min} = \min_j t_j \geq 0$ is the smallest ordering parameter. Using this inequality we find an upper bound for the interval length:

$$\mathcal{R}(W_A^{(-t)}) \leq \frac{2}{t_{\min} + 1} \prod_{j=1}^m \frac{2}{t_j + 1} \leq \left(\frac{2}{t_{\min} + 1} \right)^{m+1}. \quad (24)$$

This bound is $\mathcal{R}(W_A^{(-t)}) \leq 2^{m+1}$ for the Wigner function, but for the Husimi function becomes

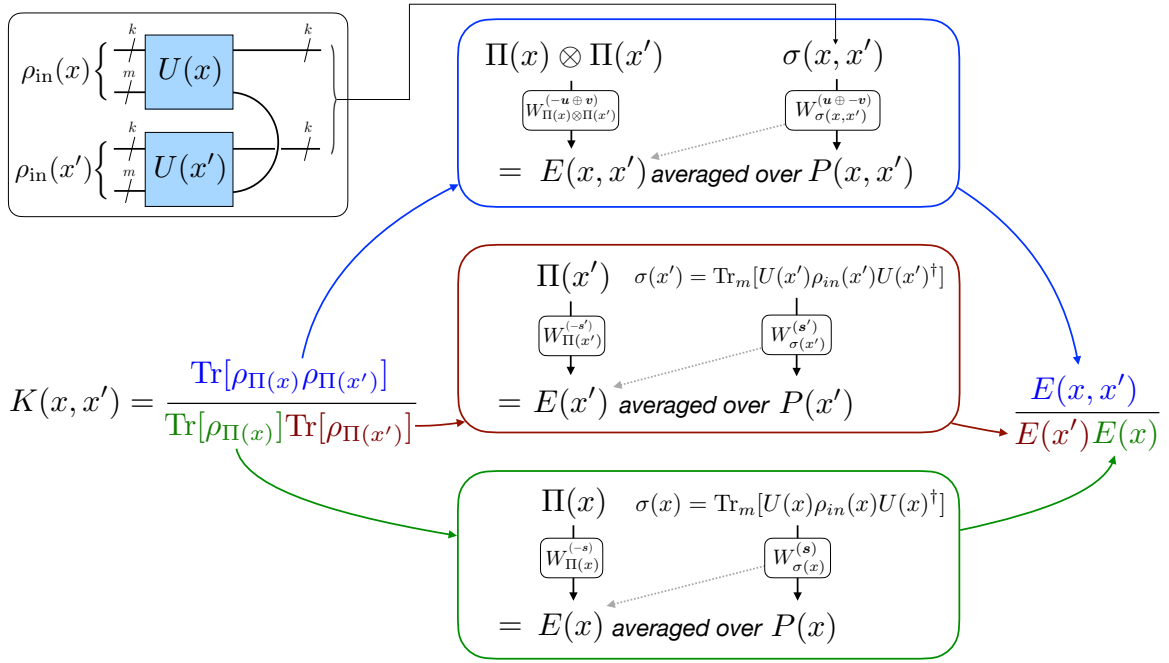


Figure 3: Classical estimation of quantum state overlap, following Algorithm 2. The algorithm involves three separate estimations, each being similar to Algorithm 1 (see Fig. 2). The dashed arrows indicate that the negativity information of the corresponding $(\mathbf{s})$ -PQDs $W_{\sigma(x)}^{(\mathbf{s})}$, $W_{\sigma(x')}^{(\mathbf{s})}$ and $W_{\sigma(x, x')}^{(\mathbf{u} \oplus -\mathbf{v})}$ are also being used to define the corresponding estimator. The k -mode states $\sigma(x)$ and $\sigma(x')$ involved in the estimation of the denominator are defined as the partial trace of the last m modes of the $(k + m)$ -mode states $U(x)\rho_{\text{in}}(x)U(x)^\dagger$ and $U(x')\rho_{\text{in}}(x')U(x')^\dagger$, respectively, while the $(2k)$ -mode state $\sigma(x, x')$ involved in the estimation of the numerator is defined as the partial overlap of the last m modes of the $(k + m)$ -mode states $U(x)\rho_{\text{in}}(x)U(x)^\dagger$ and $U(x')\rho_{\text{in}}(x')U(x')^\dagger$, as shown in the top-left circuit picture, where a partial transpose is omitted for brevity.

$\mathcal{R}(W_A^{(-I)}) \leq 1$. Therefore, for classical states with $\mathcal{N}(W_\rho^{(\mathbf{t})}) = 1$, an optimal choice for the ordering parameters is $\mathbf{t} = \mathbf{I}$ that gives $\mathcal{R}(E) \leq 1$. This implies in particular that, using Algorithm 1, the kernel function for classical states can be estimated efficiently classically.

For Algorithm 2, the first three steps are instances of Algorithm 1 and their efficiency is subject to the previous considerations. Furthermore, we show in Appendix B that for any quantum-efficient encoding, combining three independent estimates of $\text{Tr}[\rho_{\Pi(x)}]$, $\text{Tr}[\rho_{\Pi(x')}]$ and $\text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]$ that are inverse-polynomially precise with exponentially small probability of failure provides an estimate of $K(x, x') = \text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}] / (\text{Tr}[\rho_{\Pi(x)}]\text{Tr}[\rho_{\Pi(x')}])$ that is also inverse-polynomially precise with exponentially small probability of failure, which shows the efficiency of the last step of Algorithm 2.

5 Examples

In this section, we apply our previous findings to various examples. Firstly, we illustrate the use of Algorithm 1 in Section 5.1 for estimating quantum kernels based on output states of linear optical networks. Secondly, in Section 5.2, we show that there are cases where direct classical computation of the quantum kernel is efficient, such as for kernels based on Gaussian states. Finally, we illustrate the use of Algorithm 2 in Section 5.3 for a large class of quantum kernels based on non-Gaussian states obtained through the measurement of a subset of modes of Gaussian states, including adaptive measurement strategies.

5.1 Output states of linear optical networks

Linear optical networks (LON) are of particular interest because they can be simply realized, using passive optical elements such as beam split-

ters and phase shifters, yet they provide the underlying transformations in boson sampling problems [9] that are believed to be classically hard to simulate. In boson sampling, one generates samples from the output probability distribution at the output of an LON, when single-photons are injected to the input. It is interesting to note, however, that output probabilities in boson sampling can be efficiently estimated classically [9, 34]. Also, classical algorithms have been recently proposed to approximate output probabilities of boson sampling and Gaussian boson sampling using the (s) -PQDs [35].

Here, we consider a class of data-encoding states that are prepared by using LONs. In an ideal situation, a lossless LON is described by a unitary transformation $U_{\text{LON}}^\dagger(x)$ that can be defined by its action on the m -mode displacement operator, $U_{\text{LON}}^\dagger(x)D(\xi)U_{\text{LON}}(x) = D(\xi V(x))$, where $V(x)$ is an $m \times m$ unitary transfer matrix associated with the LON. Using this relation in Eqs. (8) and (9), one can find the (s) -PQD of the output state of the LON $\rho(x) = U_{\text{LON}}(x)\rho_{\text{in}}U_{\text{LON}}^\dagger(x)$,

$$W_{\rho(x)}^{(s)}(\alpha) = W_{\rho_{\text{in}}}^{(V^\dagger(x)sV(x))}(\alpha V(x)), \quad (25)$$

where $W_{\rho_{\text{in}}}^{(s)}(\alpha)$ is the (s) -PQD of the input state ρ_{in} . This relation shows that, given the (s) -PQD of the input state, one can efficiently compute the (s) -PQD of the data-encoding states $\rho(x)$, and use Algorithm 1 to estimate the kernel function.

In practice, however, any LON is lossy and hence cannot be described by a unitary transfer matrix. An interesting feature of our formalism is that it provides a practical way to take losses into account, as common sources of error, and check their effects on the negativity of the PQDs, which is directly related to the efficiency of our kernel estimation algorithms. In general, any lossy LON can be modeled as a quantum operation consisting of single-mode loss channels sandwiched between two lossless LONs [36]

$$\rho(x) = U_{\text{LON}}(x)\Lambda_\eta(\tilde{U}_{\text{LON}}(x)\rho_{\text{in}}\tilde{U}_{\text{LON}}^\dagger(x))U_{\text{LON}}^\dagger(x). \quad (26)$$

Here, $\Lambda_\eta = \Lambda_{\eta_1} \otimes \cdots \otimes \Lambda_{\eta_m}$ is an m -mode loss channel with parameters $0 \leq \eta_j \leq 1$. In Appendix C, we have derived the relation between the (s) -PQDs of the output and input states of loss channels

$$W_{\Lambda_\eta(\rho)}^{(s)}(\alpha) = \frac{1}{\det \eta} W_\rho^{(t)}(\alpha \eta^{-1/2}), \quad (27)$$

where $I - s = \eta(I - t)$ and $\eta = \text{diag}(\eta_1, \dots, \eta_m)$ in the case of diagonal matrices of ordering parameters. This relation, together with Eq. (25) enable us to investigate the effect of losses on the negative volume and range of the PQDs of data-encoding states (26), and check when efficient estimation of the kernel function is possible. Note that an alternative approach presented in [15] is to express the PQD of the state (26) in terms of the PQD of the input state and the Gaussian transition function associated with the lossy LON (see Eq.(39) of Appendix A). However, we emphasize that our classical estimation algorithms are inherently different from the classical sampling algorithms in [15], despite the similarities between the formalisms. For the sampling algorithms to work, all the PQDs must be non-negative, while our estimation algorithms always work and can be efficient even in the presence of negativity.

To further illustrate our formalism, let us assume that our initial state is product $\rho_{\text{in}} = \bigotimes_{k=1}^m \rho_k$ and losses can be modeled in terms of single-mode loss channels at the input of lossless LONs. We then consider examples of single-photon states and cat states as the initial states. Under these assumptions, Eq. (26) becomes $\rho(x) = U_{\text{LON}}(x) \bigotimes_{k=1}^m \tilde{\rho}_k U_{\text{LON}}^\dagger(x)$ where $\tilde{\rho}_k = \Lambda_{\eta_k}(\rho_k)$. Note that losses in state preparation can also be incorporated into these loss channels at the input of LONs. In this scenario, assuming that $s = sI$ in Eq. (25), the (s) -PQD of the state $\rho(x)$ can be written as

$$W_{\rho(x)}^{(sI)}(\alpha) = \prod_{k=1}^m W_{\tilde{\rho}_k}^{(s)}\left(\sum_{j=1}^m \alpha_j V_{jk}(x)\right), \quad (28)$$

where $W_{\tilde{\rho}_k}^{(s)}(\beta_k)$ is the (s) -PQD of the state injected into the k th input port of the lossless LON. In this case, the (s) -PQDs can always be efficiently computed and we can use our Algorithm 1 to estimate the kernel function

$$K(x, x') = \pi^m \int d^{2m} \alpha W_{\rho(x)}^{(sI)}(\alpha) W_{\rho(x')}^{(-sI)}(\alpha).$$

By using Eq. (28) we can compute the probability distribution in Eq. (17):

$$P(\alpha) = \prod_{k=1}^m \frac{1}{\mathcal{N}(W_{\tilde{\rho}_k}^{(s)})} \left| W_{\tilde{\rho}_k}^{(s)}\left(\sum_{j=1}^m \alpha_j V_{jk}(x)\right) \right|. \quad (29)$$

Here, $\mathcal{N}(W_{\tilde{\rho}_k}^{(s)}) = \int d^2\beta_k |W_{\tilde{\rho}_k}^{(s)}(\beta_k)|$. Notice that $\mathcal{N}(W_{\rho(x)}^{(sI)}) = \prod_{k=1}^m \mathcal{N}(W_{\tilde{\rho}_k}^{(s)})$ since lossless LONs preserve the negative volume and hence the non-classicality of quantum states [22]. We can also efficiently generate samples from the probability distribution (29) by first sampling the components of N complex vectors $\beta_1, \dots, \beta_N$ from individual probability distributions $|W_{\tilde{\rho}_k}^{(s)}(\beta_k)|/\mathcal{N}(W_{\tilde{\rho}_k}^{(s)})$, $k = 1, \dots, N$, and then calculating $\alpha_j = \beta_j V^\dagger(x)$. Given these samples, we then obtain the corresponding values for the (s) -PQD of $\rho(x')$, $W_{\rho_{\text{in}}}^{(-sI)}(\alpha_j V(x'))$, and through Eq. (18) calculate the sample mean $\frac{1}{N} \sum_{j=1}^N E(\alpha_j)$.

The sampling complexity for this class of states is described by

$$\begin{aligned} \mathcal{R}(E) &\leq 2\pi^m \mathcal{N}(W_{\rho(x)}^{(sI)}) \max_{\alpha} |W_{\rho(x')}^{(-sI)}(\alpha)| \\ &= 2 \prod_{k=1}^m \left(\pi \mathcal{N}(W_{\tilde{\rho}_k}^{(s)}) \max_{\alpha} |W_{\tilde{\rho}_k}^{(-s)}(\alpha)| \right), \end{aligned} \quad (30)$$

which is completely independent of the data-encoding LON unitary operations and depends only on the (s) -PQD of the input states. Therefore, for a given input state $\rho_{\text{in}} = \bigotimes_{k=1}^m \rho_k$ and loss parameters, one can minimize $\mathcal{R}(E)$ by finding an optimal ordering parameter s and determine the scaling for the number of required samples. Notice, however, that one could use a more optimal estimation procedure with different ordering parameters for each input mode, in which case this would depend on the LON description. Using Eq. (30), it is easy to verify that if all states $\tilde{\rho}_k$ are classical, then for the optimal choice of $s = 1$ we have $\mathcal{R}(E) = 1$.

5.1.1 Example 1: Single-photon states

Let us consider the case where all the input states are single-photon states $\rho_k = |1\rangle\langle 1|$ in the above formalism. Hence, the input states of lossless LONs are $\tilde{\rho}_k = \Lambda_{\eta_k}(|1\rangle\langle 1|) = (1 - \eta_k)|0\rangle\langle 0| + \eta_k|1\rangle\langle 1|$. Using the single-mode version of Eq. (9), the (s) -PQD of $\tilde{\rho}_k$ is given by

$$W_{\tilde{\rho}_k}^{(s)}(\alpha) = \frac{2(1-s)(1-s-2\eta_k) + 8\eta_k|\alpha|^2}{\pi(1-s)^3} e^{-2|\alpha|^2/(1-s)}$$

which is non-negative if $s \leq 1 - 2\eta_k$. Moreover, if $s > 1 - 2\eta_k$, the negative volume is given by

$$\mathcal{N}(W_{\tilde{\rho}_k}^{(s)}) = \frac{4\eta_k}{1-s} \exp\left(\frac{1-s-2\eta_k}{2\eta_k}\right) - 1. \quad (31)$$

To examine the bound (30), we should also consider the maximum value of $W_{\tilde{\rho}_k}^{(-s)}(\alpha)$. This function has two extremal values at $|\alpha_0|^2 = 0$ and $|\alpha_1|^2 = (1+s)(4\eta_k - 1 - s)/(2\eta_k)$, which are given by

$$W_{\tilde{\rho}_k}^{(s)}(0) = \frac{2(1+s-2\eta_k)}{\pi(1+s)^2},$$

and

$$W_{\tilde{\rho}_k}^{(s)}(\alpha_1) = \frac{4\eta}{\pi(1+s)^2} \exp\left(\frac{s+1-4\eta_k}{2\eta_k}\right).$$

Therefore, using these expressions and Eq. (31) in Eq. (30), and then optimizing over s , we can find how the sample complexity scales with the number of modes.

As an example, assuming that $\eta_k = 1/2$ for all k , we can verify that $s = 0$ is an optimal ordering parameter since $\mathcal{N}_{\rho_k}^{(0)} = 1$ and $0 \leq W_{\rho_k}^{(0)}(\alpha) \leq 2/(e\pi)$. Therefore, in this case, $\mathcal{R}(E) \leq 1$ and the kernel function can be estimated efficiently to within an additive error $\epsilon = 1/\text{poly}(m)$ with an exponentially small probability of failure. Numerical analysis can be utilized to handle other values of η_k . For instance, by Eq. (30), it can be seen that for $\eta = 0.85$ and the ordering parameter $s = 0.3$, $\mathcal{R}(E) \leq 2$, independent of the number of modes, and hence our estimation algorithm is still efficient.

Note that due to their generality, phase-space methods may be outperformed by other classical simulation techniques. For instance, we derive a variant of Gurvits' algorithm for estimating the permanent [34] in Appendix D which allows us to perform efficient classical estimation of quantum kernels based on lossy photonic states for *any* loss parameter.

5.1.2 Example 2: Cat states

Another class of states that we consider as the input states for LONs are the cat states

$$|\text{cat}_k\rangle = \frac{|\gamma_k\rangle + |-\gamma_k\rangle}{\sqrt{2 + 2e^{-2|\gamma_k|^2}}}, \quad (32)$$

where $|\gamma_k\rangle$ denotes a coherent state. By using Eq. (8), the (s) -PQD of this state is given by

$$W_{|\text{cat}_k\rangle}^{(s)}(\alpha) = \frac{1}{\pi(1-s)(1+e^{-2|\gamma_k|^2})} \left(e^{-2\frac{|\alpha+\gamma_k|^2}{1-s}} + e^{-2\frac{|\alpha-\gamma_k|^2}{1-s}} + 2e^{-2|\gamma_k|^2} \Re \left(e^{-2\frac{(\alpha+\gamma_k)(\alpha^*-\gamma_k^*)}{1-s}} \right) \right), \quad (33)$$

where $\Re$ denotes the real part of the expression. Using this equation and Eq. (27), we can then compute the (s) -PQD of the states after loss channels, $\tilde{\rho}_k = \Lambda_{\eta_k}(|\text{cat}_k\rangle\langle\text{cat}_k|)$,

$$W_{\tilde{\rho}_k}^{(s)}(\alpha) = \frac{1}{\pi(1-s)(1+e^{-2|\gamma_k|^2})} \left(e^{-2\frac{|\alpha+\sqrt{\eta_k}\gamma_k|^2}{1-s}} + e^{-2\frac{|\alpha-\sqrt{\eta_k}\gamma_k|^2}{1-s}} + 2e^{-2|\gamma_k|^2} \Re \left(e^{-2\frac{(\alpha+\sqrt{\eta_k}\gamma_k)(\alpha^*-\sqrt{\eta_k}\gamma_k^*)}{1-s}} \right) \right). \quad (34)$$

Given the parameters for input cat states and losses, one can compute the upper bound on the range of the estimator $\mathcal{R}(E)$ in Eq. (30), and find an optimal ordering parameter s by using this expression.

For example, numerical analysis shows that if $\gamma_k = 4$ and $\eta_k = 0.8$ for all k , then $\pi\mathcal{N}(W_{\tilde{\rho}_k}^{(s)}) \max_{\alpha} |W_{\tilde{\rho}_k}^{(-s)}(\alpha)| < 1$ for $s = 0.1$. Therefore, in this case, we have $\mathcal{R}(E) < 2$, independent of the number of modes, and our Algorithm 1 can be used to estimate the kernel function efficiently.

5.2 Gaussian states

We now focus on Gaussian states. Choosing the parameter $\mathbf{s} = \mathbf{0}$ provides an efficient classical estimation of the quantum kernel through Algorithm 1. However, instead of using the estimation algorithm we described, we can check whether the kernel function (10) can be computed directly using conventional methods for computing integrals. Indeed, for Gaussian states the corresponding quantum kernel functions can be computed exactly analytically.

Gaussian states have Gaussian Wigner functions that can be described in terms of the mean values $\bar{\mathbf{r}} = \text{Tr}[\rho \mathbf{r}]$, where $\mathbf{r} = (q_1, p_1, \dots, q_m, p_m)^T$ is the vector of canonical operators $q_j = (a_j + a_j^\dagger)/\sqrt{2}$ and $p_j = i(a_j^\dagger - a_j)/\sqrt{2}$,

and the covariance matrix $\Sigma_{jk} = \text{Tr}[\rho(\mathbf{r}_j \mathbf{r}_k + \mathbf{r}_k \mathbf{r}_j)]/2 - \bar{\mathbf{r}}_j \bar{\mathbf{r}}_k$. Indeed, for such a Gaussian state ρ we have [37]:

$$W_{\rho}^{(s)}(\alpha) = \frac{e^{-\frac{1}{2}(\alpha - \bar{\mathbf{r}})(\Sigma - \mathbf{s} \oplus \mathbf{s})^{-1}(\alpha - \bar{\mathbf{r}})^T}}{(2\pi)^m \sqrt{\det(\Sigma - \mathbf{s} \oplus \mathbf{s})}}, \quad (35)$$

for all $\mathbf{s}$ such that $\Sigma - \mathbf{s} \oplus \mathbf{s}$ is positive definite, where we used that $(\mathbf{s})$ -PQDs are related to the Wigner function by a Gaussian convolution. For single-mode Gaussian states, the minimal valid choice for $\tau := \frac{1}{2}(1-s) \in [0, \frac{1}{2}]$ is known as the non-classical depth [23]. This definition readily extends to the multimode setting (we use a simplified version of the general definition in [24]):

Definition 1 (Non-classical depth). *The non-classical depth of a quantum state ρ is the minimal value $\tau = \frac{1}{2}(1-s) \in [0, 1]$ such that the $(\mathbf{s})$ -PQD of the state ρ is non-negative for $\mathbf{s} = \mathbf{sI}$.*

By Eq. (35), for multimode Gaussian states the minimal eigenvalue of the covariance matrix encodes this information. Using Eq. (10) for $\mathbf{s} = \mathbf{0}$ one can verify that the kernel function is given by

$$K(x, x') = \frac{e^{-\frac{1}{2}(\bar{\mathbf{r}} - \bar{\mathbf{r}}')(\Sigma(x) + \Sigma(x'))^{-1}(\bar{\mathbf{r}} - \bar{\mathbf{r}}')^T}}{2^{-m} \sqrt{\det(\Sigma(x) + \Sigma(x'))}}.$$

Thus, for Gaussian states we do not really need our Monte Carlo-based method to estimate the kernel function, and quantum machine learning protocols that use data-encoding Gaussian states can be efficiently simulated by classical algorithms [38]. We note that it is strongly believed that sampling from the photon-counting probability distributions for Gaussian states cannot be simulated efficiently classically [39–41]. Thus, although computing the kernel function for such states is easy, sampling from their photon-counting distribution is hard.

5.3 Partially measured Gaussian states

Given the limitations of Gaussian states for quantum kernel methods, we can ask whether non-Gaussian states can be helpful. A standard way to engineer a non-Gaussian state is to perform non-Gaussian measurements on a subset of the modes of a Gaussian state (see [42] and references therein). We thus consider the special case

of Eq. (1) consisting of quantum states $\rho(x)$ prepared by measuring some of the output modes of a multimode Gaussian state, i.e.,

$$x \mapsto \rho(x) = \frac{\text{Tr}_k[(\Pi(x) \otimes \mathbb{I}_m)\rho_G(x)]}{\text{Tr}[(\Pi(x) \otimes \mathbb{I}_m)\rho_G(x)]}, \quad (36)$$

where $\rho_G(x) = U(x)\rho_{\text{in}}(x)U(x)^\dagger$ is an $(k+m)$ -mode Gaussian state and $\Pi(x) = \bigotimes_{j=1}^k \Pi_j(x)$ is a tensor product of (possibly non-Gaussian) POVM elements. Recall that a quantum-efficient encoding refers to the fact that such states may be efficiently prepared using a quantum computer, a property which can be summarized here by $\text{Tr}[(\Pi(x) \otimes \mathbb{I}_m)\rho_G(x)] \geq 1/(\text{poly}(m))$.

Using our kernel estimation formalism and Algorithm 2 in particular, we show that, when kernel estimation is quantum-efficient, classical kernel estimation is also efficient whenever either the number of measured modes or the non-classicality of the Gaussian states involved is too small:

Theorem 1. *For any classical data x , let $\rho(x)$ be a quantum state encoding over m modes obtained by performing a measurement of the first k modes of a $(k+m)$ -mode Gaussian state $\rho_G(x)$, as in Eq. (36). Let $\tau(x)$ denote the nonclassical depth of $\rho_G(x)$ (see Definition 1) and let $\tau(x, x') = \max(\tau(x), \tau(x')) \in [0, \frac{1}{2}]$. Then, assuming that the encoding is quantum-efficient, Algorithm 2 provides an estimate of the quantum kernel $K(x, x') = \text{Tr}[\rho(x)\rho(x')]$ with additive precision ϵ and success probability $1 - \delta$ in time*

$$O\left(\frac{\log(\frac{2}{\delta})}{\epsilon^2(1 - \tau(x, x'))^{4k+2}} \text{poly}(m)\right), \quad (37)$$

In particular, this provides an efficient classical algorithm for quantum kernel estimation whenever $k = O(\log m)$ or else $\tau(x, x') = O(\log m/k)$.

We give a proof of this theorem in Appendix E, which combines a careful analysis of the time complexity of Algorithm 2 together with new properties of the non-classical depth of Gaussian states.

A nontrivial consequence of Theorem 1 is that the efficiency of the classical simulation is independent of the non-Gaussianity of the measurements: even though these can inject a lot of negativity in the prepared state, as measured by the negative volume, this negativity does not affect the classical simulability through Algorithm 2,

because the POVM elements only contribute to defining the estimators in Algorithm 2 and their non-Gaussianity does not substantially change the range of these estimators. In particular, even when making very non-Gaussian measurements (such as detecting many photons), classical simulation is always efficient as long as only a few modes $k = O(\log m)$ are measured.

What about when the number of measured modes is larger? There, Theorem 1 shows that classical simulation is still efficient when the Gaussian state being measured has small non-classicality, as quantified by the non-classical depth (see Definition 1). This notion of non-classicality is directly related to the amount of thermal noise necessary to make the state fully classical [23], i.e., with non-negative P function, and for a Gaussian state it is related to the minimum eigenvalue of its covariance matrix, see Eq. (35). This quantity can in turn be bounded by a function of the squeezing parameters and the symplectic eigenvalues encoding the impurity of the corresponding Gaussian state.

For illustration purposes, let us consider $\rho_G := U\rho_{\text{in}}U^\dagger$ with ρ_{in} being a tensor product of identical single-mode thermal states ν , and U being a Gaussian unitary operator. By virtue of the Euler (or Bloch–Messiah) decomposition [37] we may write $U = DVS V'$ with D being a tensor product of single-mode displacement operators, S being a tensor product of single-mode squeezing operators, and V, V' being passive linear transforms describing the action of lossless LONs. Hence, $\rho_G = DVS\nu^{\otimes(k+m)}S^\dagger V'^\dagger S^\dagger V^\dagger D^\dagger = DVS\nu^{\otimes(k+m)}S^\dagger V^\dagger D^\dagger$, where we used the fact that a tensor product of identical single-mode thermal states is invariant under lossless LONs. This state has the following covariance matrix [37]:

$$O_V \begin{pmatrix} \frac{1}{\mu}\Delta^2 & 0 \\ 0 & \frac{1}{\mu}\Delta^{-2} \end{pmatrix} O_V^\top, \quad (38)$$

where O_V is the symplectic orthogonal matrix associated to the LON V , Δ is a diagonal matrix containing the squeezing parameters and μ is the purity of the single-mode thermal state ν . Writing $\frac{1}{r}$ for the minimal squeezing parameter smaller than 1, with $r \geq 1$, the non-classical depth of the state is given by $\tau = \frac{1}{2}(1 - \frac{1}{r^2\mu})$ by Eq. (35), and the condition $\tau = O(\log m/k)$ in Theorem 1 implies that classical estimation

of quantum kernels based on partially measured Gaussian states of the form of ρ_G is efficient whenever the squeezing r or the purity μ are too small.

Theorem 1 also allows us to investigate the effect of lossy state preparation: with Eq. (27) and Definition 1, uniform losses of transmissivity η over all modes map the non-classical depth from τ to $\eta\tau$, in which case the classical estimation provided by Algorithm 2 is efficient whenever $k = O(\log m)$ or else $\eta\tau(x, x') = O(\log m/k)$.

Finally, up to taking mixtures, our results on partially measured Gaussian states also cover the case of quantum states prepared by Gaussian computations, together with *adaptive* measurements, i.e., intermediate measurements whose outcome can drive the rest of the computation. In particular, we show that classical estimation of the corresponding quantum kernel functions is efficient under the conditions of Theorem 1, if the number of possible adaptive measurement outcomes is small enough (see Appendix F for details).

6 Conclusion and outlook

We have introduced a framework based on phase-space quasi-probability distributions for the classical estimation of quantum kernel functions in machine learning. Our sufficient conditions for efficient classical simulation are based on negative volume, non-classical depth, and excess range of quasi-probability distributions, and identify phase-space negativity as an essential resource for achieving computational advantages in quantum machine learning with kernel methods. Our formalism can also be used to investigate the effect of errors and imperfections in quantum machine learning devices by examining their impact on the negativity of quasi-probability distributions.

By considering various examples based on variants of the boson sampling model, we have showcased how sampling from the output probability distribution of a quantum circuit can be classically hard, yet supervised machine learning using the same circuit can be classically efficient.

Moreover, we have identified a subtle interplay between the quantum computational resources at hand: if no phase-space negativity is involved, and in the case of Gaussian measurements in particular, quantum kernels based on partially

measured Gaussian states can be efficiently estimated classically using our Algorithm 1 (see Section 5.2). When phase-space negativity is present in the measurements, then our Algorithm 2 still allows for efficient classical estimation of the corresponding quantum kernels as long as the number of measured modes is small enough or the non-classical depth of the Gaussian states involved is small enough (see Section 5.3). In other terms, quantum computational advantage for quantum kernel estimation is only possible in this setting by combining Gaussian non-classical resources (squeezing) and non-Gaussian resources (phase-space negative volume). A similar situation arises in the context of Gaussian boson sampling [43], where squeezing is a necessary ingredient together with non-Gaussianity for quantum computational advantage through sampling [44, 45].

Our results could be extended in a few directions. The sample complexities of our classical simulation algorithms are naturally expressed using non-classical measures relating to phase-space negativity; it would be interesting to relate these measures to other existing ones such as quadrature coherence scale [46], which provides an estimation of the distance to the set of classical states, or stellar rank [47], according to which a classification of bosonic kernels for quantum machine learning was recently derived [48]. Another direction could be to use our framework to analyze in more details how specific imperfections in implementations of quantum kernel methods, and the SWAP test in particular [49–51], could ease classical simulability. One could also apply the presented framework to the case of non-linear optical approaches such as the optical Ising machine [52] or Kerr-based kernels [53]. Moreover, it would be interesting to generalize the presented approach to the case of discrete-variable quasi-probability distributions using frame theory [54, 55].

Acknowledgements. U.C. acknowledges inspiring discussions with M. Walschaers, M. Frigerio, F. Arzani, J. Davis, M. Garnier, H. Thomas, P.E. Emeriau, S. Mehraban and D. Hangleiter. R.G. acknowledges discussion with C. Simon at the early stage of this project. U.C. acknowledges funding from the European Union’s Horizon Europe Framework Programme (EIC Pathfinder Challenge project Veriqub) under Grant Agree-

References

- [1] J. Preskill, “Quantum Computing in the NISQ era and beyond,” *Quantum* **2**, 79 (2018).
- [2] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. Brandao, D. A. Buell, *et al.*, “Quantum supremacy using a programmable superconducting processor,” *Nature* **574**, 505–510 (2019).
- [3] I. M. Georgescu, S. Ashhab, and F. Nori, “Quantum simulation,” *Rev. Mod. Phys.* **86**, 153–185 (2014).
- [4] A. J. McCaskey, Z. P. Parks, J. Jakowski, S. V. Moore, T. D. Morris, T. S. Humble, and R. C. Pooser, “Quantum chemistry as a benchmark for near-term quantum computers,” *npj Quantum Information* **5**, 99 (2019).
- [5] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature* **549**, 195–202 (2017).
- [6] E. Farhi and H. Neven, “Classification with quantum neural networks on near term processors,” [arXiv:1802.06002](https://arxiv.org/abs/1802.06002).
- [7] B. M. Terhal and D. P. DiVincenzo, “Adaptive quantum computation, constant depth quantum circuits and arthur-merlin games,” *Quantum Inf. Comput.* **4**, 134–145 (2004).
- [8] M. J. Bremner, A. Montanaro, and D. J. Shepherd, “Average-Case Complexity Versus Approximate Simulation of Commuting Quantum Computations,” *Phys. Rev. Lett.* **117**, 080501 (2016).
- [9] S. Aaronson and A. Arkhipov, “The Computational Complexity of Linear Optics,” *Theory of Computing* **9**, 143–252 (2013).
- [10] S. Bravyi, D. Gosset, and R. König, “Quantum advantage with shallow circuits,” *Science* **362**, 308–311 (2018).
- [11] K. E. Cahill and R. J. Glauber, “Density Operators and Quasiprobability Distributions,” *Phys. Rev.* **177**, 1882–1902 (1969).
- [12] R. W. Spekkens, “Negativity and Contextuality are Equivalent Notions of Nonclassicality,” *Phys. Rev. Lett.* **101**, 020401 (2008).
- [13] A. Mari and J. Eisert, “Positive Wigner Functions Render Classical Simulation of Quantum Computation Efficient,” *Phys. Rev. Lett.* **109**, 230503 (2012).
- [14] V. Veitch, N. Wiebe, C. Ferrie, and J. Emerson, “Efficient simulation scheme for a class of quantum optics experiments with non-negative Wigner representation,” *New Journal of Physics* **15**, 013037 (2013).
- [15] S. Rahimi-Keshari, T. C. Ralph, and C. M. Caves, “Sufficient Conditions for Efficient Classical Simulation of Quantum Optics,” *Phys. Rev. X* **6**, 021039 (2016).
- [16] D. Stahlke, “Quantum interference as a resource for quantum speedup,” *Phys. Rev. A* **90**, 022302 (2014).
- [17] H. Pashayan, J. J. Wallman, and S. D. Bartlett, “Estimating Outcome Probabilities of Quantum Circuits Using Quasiprobabilities,” *Phys. Rev. Lett.* **115**, 070501 (2015).
- [18] R. Mengoni and A. Di Pierro, “Kernel methods in quantum machine learning,” *Quantum Machine Intelligence* **1**, 65–71 (2019).
- [19] R. Ghobadi, “Nonclassical kernels in continuous-variable systems,” *Physical Review A* **104**, 052403 (2021).
- [20] M. Bohmann and E. Agudelo, “Phase-space inequalities beyond negativities,” *Physical Review Letters* **124**, 133601 (2020).
- [21] A. Kenfack and K. Życzkowski, “Negativity of the Wigner function as an indicator of non-classicality,” *Journal of Optics B: Quantum and Semiclassical Optics* **6**, 396 (2004).
- [22] F. Albarelli, M. G. Genoni, M. G. A. Paris, and A. Ferraro, “Resource theory of quantum non-Gaussianity and Wigner negativity,” *Phys. Rev. A* **98**, 052350 (2018).
- [23] C. T. Lee, “Measure of the nonclassicality of nonclassical states,” *Physical Review A* **44**, R2775 (1991).
- [24] K. K. Sabapathy, “Process output nonclassicality and nonclassicality depth of quantum-optical channels,” *Phys. Rev. A* **93**, 042103 (2016).

- [25] M. Schuld and F. Petruccione, *Quantum Models as Kernel Methods*, pp. 217–245. Springer International Publishing, Cham, 2021.
- [26] B. Schölkopf, R. Herbrich, and A. J. Smola, “A Generalized Representer Theorem,” in *Computational Learning Theory*, D. Helmbold and B. Williamson, eds., pp. 416–426. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.
- [27] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” *The Annals of Statistics* **36**, 1171–1220 (2008).
- [28] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, “Supervised learning with quantum-enhanced feature spaces,” *Nature* **567**, 209–212 (2019).
- [29] M. Schuld and N. Killoran, “Quantum Machine Learning in Feature Hilbert Spaces,” *Phys. Rev. Lett.* **122**, 040504 (2019).
- [30] H. Buhrman, R. Cleve, J. Watrous, and R. De Wolf, “Quantum fingerprinting,” *Physical review letters* **87**, 167902 (2001).
- [31] M. Hillery, R. O’Connell, M. Scully, and E. Wigner, “Distribution functions in physics: Fundamentals,” *Physics Reports* **106**, 121–167 (1984).
- [32] R. P. Rundell, P. W. Mills, T. Tilma, J. H. Samson, and M. J. Everitt, “Simple procedure for phase-space measurement and entanglement validation,” *Phys. Rev. A* **96**, 022117 (2017).
- [33] W. Hoeffding, “Probability Inequalities for Sums of Bounded Random Variables,” *Journal of the American Statistical Association* **58**, 13–30 (1963).
- [34] L. Gurvits, “On the Complexity of Mixed Discriminants and Related Problems,” in *Mathematical Foundations of Computer Science 2005*, J. Jędrzejowicz and A. Szepietowski, eds., pp. 447–458. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
- [35] Y. Lim and C. Oh, “Approximating outcome probabilities of linear optical circuits,” *npj Quantum Information* **9**, 124 (2023).
- [36] S. Rahimi-Keshari, S. Baghbanzadeh, and C. M. Caves, “In situ characterization of linear-optical networks in randomized boson sampling,” *Physical Review A* **101**, 043809 (2020).
- [37] A. Ferraro, S. Olivares, and M. G. A. Paris, “Gaussian States in Quantum Information,” [arxiv:quant-ph/0503237](https://arxiv.org/abs/quant-ph/0503237).
- [38] M. Schuld, K. Brádler, R. Israel, D. Su, and B. Gupt, “Measuring the similarity of graphs with a Gaussian boson sampler,” *Phys. Rev. A* **101**, 032314 (2020).
- [39] A. P. Lund, A. Laing, S. Rahimi-Keshari, T. Rudolph, J. L. O’Brien, and T. C. Ralph, “Boson Sampling from a Gaussian State,” *Phys. Rev. Lett.* **113**, 100502 (2014).
- [40] S. Rahimi-Keshari, A. P. Lund, and T. C. Ralph, “What Can Quantum Optics Say about Computational Complexity Theory?,” *Phys. Rev. Lett.* **114**, 060501 (2015).
- [41] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, “Gaussian Boson Sampling,” *Phys. Rev. Lett.* **119**, 170501 (2017).
- [42] A. Lvovsky, P. Grangier, A. Ourjoumtsev, V. Parigi, M. Sasaki, and R. Tualle-Brouiri, “Production and applications of non-Gaussian quantum states of light,” [arXiv:2006.16985](https://arxiv.org/abs/2006.16985).
- [43] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, “Gaussian boson sampling,” *Physical review letters* **119**, 170501 (2017).
- [44] U. Chabaud and S. Mehraban, “Holomorphic representation of quantum computations,” *Quantum* **6**, 831 (2022).
- [45] U. Chabaud and M. Walschaers, “Resources for bosonic quantum computational advantage,” *Physical Review Letters* **130**, 090602 (2023).
- [46] A. Hertz and S. De Bièvre, “Quadrature coherence scale driven fast decoherence of bosonic quantum field states,” *Physical Review Letters* **124**, 090402 (2020).
- [47] U. Chabaud, D. Markham, and F. Grosshans, “Stellar representation of non-Gaussian quantum states,” *Physical Review Letters* **124**, 063605 (2020).
- [48] L. J. Henderson, R. Goel, and S. Shrapnel, “Quantum kernel machine learning with continuous variables,” [arXiv:2401.05647](https://arxiv.org/abs/2401.05647).
- [49] Y. Y. Gao, B. J. Lester, Y. Zhang, C. Wang, S. Rosenblum, L. Frunzio,

- L. Jiang, S. Girvin, and R. J. Schoelkopf, “Programmable interference between two microwave quantum memories,” *Physical Review X* **8**, 021073 (2018).
- [50] H. Gan, G. Maslennikov, K.-W. Tseng, C. Nguyen, and D. Matsukevich, “Hybrid quantum computing with conditional beam splitter gate in trapped ion system,” *Physical review letters* **124**, 170502 (2020).
- [51] O. Černotík, I. Pietikäinen, S. Puri, S. Girvin, and R. Filip, “Swap-test interferometry with biased qubit noise,” *Physical Review Research* **6**, 033074 (2024).
- [52] Z. Wang, A. Marandi, K. Wen, R. L. Byer, and Y. Yamamoto, “Coherent Ising machine based on degenerate optical parametric oscillators,” *Phys. Rev. A* **88**, 063853 (2013).
- [53] S. Dehdashti, P. Tiwari, K. H. E. Safty, P. Bruza, and J. Notzel, “Enhancing Quantum Machine Learning: The Power of Non-Linear Optical Reproducing Kernels,” [arXiv:2407.13809](https://arxiv.org/abs/2407.13809).
- [54] C. Ferrie and J. Emerson, “Frame representations of quantum mechanics and the necessity of negativity in quasi-probability representations,” *Journal of Physics A: Mathematical and Theoretical* **41**, 352001 (2008).
- [55] C. Ferrie and J. Emerson, “Framed Hilbert space: hanging the quasi-probability pictures of quantum theory,” *New Journal of Physics* **11**, 063040 (2009).
- [56] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, “Gaussian quantum information,” *Reviews of Modern Physics* **84**, 621 (2012).
- [57] S. Boyd, S. P. Boyd, and L. Vandenberghe, “Convex optimization,” *Cambridge university press*, 2004.
- [58] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, “Introduction to algorithms,” *MIT press*, 2009.
- [59] H. J. Briegel, D. E. Browne, W. Dür, R. Raussendorf, and M. Van den Nest, “Measurement-based quantum computation,” *Nature Physics* **5**, 19–26 (2009).
- [60] E. Knill, R. Laflamme, and G. J. Milburn, “A scheme for efficient quantum computation with linear optics,” *nature* **409**, 46–52 (2001).
- [61] S. Bartolucci, P. Birchall, H. Bombin, H. Cable, C. Dawson, M. Gimeno-Segovia, E. Johnston, K. Kieling, N. Nickerson, M. Pant, *et al.*, “Fusion-based quantum computation,” *Nature Communications* **14**, 912 (2023).
- [62] U. Chabaud, D. Markham, and A. Sohbi, “Quantum machine learning with adaptive linear optics,” *Quantum* **5**, 496 (2021).
- [63] D. J. Brod, E. F. Galvão, A. Crespi, R. Osellame, N. Spagnolo, and F. Sciarrino, “Photonic implementation of boson sampling: a review,” *Advanced Photonics* **1**, 034001–034001 (2019).
- [64] H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, *et al.*, “Quantum computational advantage using photons,” *Science* **370**, 1460–1463 (2020).

Appendix

A Computation of phase-space quasi-probability distributions

We present a general method in order to describe the $(\mathbf{s})$ -PQDs when the states of interest $\rho(x)$ are obtained by applying a unitary circuit $U(x)$ to an initial state $\rho_{\text{in}}(x)$. We further assume that the initial state $\rho_{\text{in}}(x) = \rho_1(x) \otimes \cdots \otimes \rho_m(x)$ is a product state, in which case its $(\mathbf{s})$ -PQD can be efficiently described. Then, the $(\mathbf{s})$ -PQD of the state $\rho(x) = U(x)\rho_{\text{in}}(x)U(x)^\dagger$ is given by

$$W_{\rho(x)}^{(\mathbf{s})}(\boldsymbol{\alpha}) = \int_{\mathbb{C}^m} d^{2m}\boldsymbol{\beta} W_{\rho_{\text{in}}(x)}^{(\mathbf{t})}(\boldsymbol{\beta}) T_{U(x)}^{(\mathbf{s}, -\mathbf{t})}(\boldsymbol{\alpha}|\boldsymbol{\beta}), \quad (39)$$

where $T_{U(x)}^{(\mathbf{s}, -\mathbf{t})}(\boldsymbol{\alpha}|\boldsymbol{\beta}) = \pi^m \text{Tr}[U(x)\Delta^{(-\mathbf{t})}(\boldsymbol{\alpha})U^\dagger(x)\Delta^{(\mathbf{s})}(\boldsymbol{\alpha})]$ is the transition function [15] associated to $U(x)$, and $W_{\rho_{\text{in}}(x)}^{(\mathbf{t})}(\boldsymbol{\beta})$ is the PQD of the initial product state given by $W_{\rho_{\text{in}}(x)}^{(\mathbf{t})}(\boldsymbol{\beta}) = \prod_{j=1}^m W_{\rho_j(x)}^{(\mathbf{t}_j)}(\beta_j)$.

We note that in certain cases such as Gaussian circuits, the transition function $T_{U(x)}^{(\mathbf{s}, -\mathbf{t})}(\boldsymbol{\alpha}|\boldsymbol{\beta})$ can be efficiently computed. More generally, one can decompose the encoding circuit $U(x) = u_L(x) \cdots u_1(x)$ into L layers of unitaries acting on at most a constant number of modes, in which case we have

$$T_{U(x)}^{(\mathbf{r}_L, -\mathbf{r}_0)}(\boldsymbol{\gamma}_L|\boldsymbol{\gamma}_0) = \int_{(\mathbb{C}^m)^{L-1}} d^{2m}\boldsymbol{\gamma}_1 \cdots d^{2m}\boldsymbol{\gamma}_{L-1} \prod_{l=1}^L T_{u_l(x)}^{(\mathbf{r}_l, -\mathbf{r}_{l-1})}(\boldsymbol{\gamma}_l|\boldsymbol{\gamma}_{l-1}). \quad (40)$$

Therefore, using Eqs. (39) and (40), the kernel function can be expressed in terms of functions that can be computed efficiently and used to draw samples from the distribution (17), as we show below.

Writing the kernel function as $K(x, x') = \text{Tr}[U^\dagger(x')U(x)\rho U^\dagger(x)U(x')\rho]$ and using $T_{U^\dagger(x)}^{(\mathbf{s}, -\mathbf{t})}(\boldsymbol{\alpha}|\boldsymbol{\beta}) = T_{U(x)}^{(\mathbf{s}, -\mathbf{t})}(\boldsymbol{\beta}|\boldsymbol{\alpha})$, one can define the probability distribution

$$P(\vec{\gamma}) = \frac{1}{\mathcal{N}} |W_{\rho(x)}^{(\mathbf{r}_0)}(\boldsymbol{\gamma}_0)| \prod_{l=1}^L |T_{u_l(x)}^{(\mathbf{r}_l, -\mathbf{r}_{l-1})}(\boldsymbol{\gamma}_l|\boldsymbol{\gamma}_{l-1})| \prod_{k=L+1}^{2L} |T_{u_{k-L}(x')}^{(\mathbf{r}_k, -\mathbf{r}_{k-1})}(\boldsymbol{\gamma}_k|\boldsymbol{\gamma}_{k-1})|, \quad (41)$$

where $\mathcal{N} = \mathcal{N}(W_{\rho(x)}^{(\mathbf{r}_0)}) \prod_{l=1}^L \mathcal{N}(T_{u_l(x)}^{(\mathbf{r}_l, -\mathbf{r}_{l-1})}) \prod_{k=L+1}^{2L} \mathcal{N}(T_{u_{k-L}(x')}^{(\mathbf{r}_k, -\mathbf{r}_{k-1})})$ is the total negative volume, and $\vec{\gamma} = (\boldsymbol{\gamma}_0, \dots, \boldsymbol{\gamma}_{2L})$ is the vector of $(2L+1)m$ complex numbers. Viewing this expression as a Markov chain, by sampling from the distribution $|W_{\rho(x)}^{(\mathbf{r}_0)}(\boldsymbol{\gamma}_0)|/\mathcal{N}(W_{\rho(x)}^{(\mathbf{r}_0)})$ associated to the initial state that is known to be product, as well as other conditional probability distributions associated to the transition functions, one can generate N random samples $\vec{\gamma}_1, \dots, \vec{\gamma}_N$. Then, using the estimator

$$E(\vec{\gamma}) := \mathcal{N}_{\rho}^{(\mathbf{r}_0)} \text{sgn}[W_{\rho(x')}^{(\mathbf{r}_0)}(\boldsymbol{\gamma}_0)] \pi^m W_{\rho(x')}^{(\mathbf{r}_{2L})}(\boldsymbol{\gamma}_{2L}) \prod_{l=1}^L \mathcal{N}_{u_l(x)}^{(\mathbf{r}_l)} \text{sgn}[T_{u_l(x)}^{(\mathbf{r}_l, -\mathbf{r}_{l-1})}(\boldsymbol{\gamma}_l|\boldsymbol{\gamma}_{l-1})] \\ \times \prod_{k=L+1}^{2L} \mathcal{N}_{u_{k-L}(x')}^{(\mathbf{r}_k)} \text{sgn}[T_{u_{k-L}(x')}^{(\mathbf{r}_k, -\mathbf{r}_{k-1})}(\boldsymbol{\gamma}_k|\boldsymbol{\gamma}_{k-1})], \quad (42)$$

the sample mean $\frac{1}{N} \sum_j E(\vec{\gamma}_j)$ can be computed. As discussed in the main text, the relation between the estimation error and the probability of failure is given by Hoeffding's inequality. Therefore, following a similar argument, the kernel function can be estimated to within error $\epsilon = 1/\text{poly}(m)$ and an exponentially small probability of failure, if one can find ordering parameters $\{\mathbf{r}_k\}$ such that $\pi^m \mathcal{N} \times [\max_{\boldsymbol{\gamma}_b} W_{\rho(x')}^{(\mathbf{r}_{2L})}(\boldsymbol{\gamma}_b) - \min_{\boldsymbol{\gamma}_a} W_{\rho(x')}^{(\mathbf{r}_{2L})}(\boldsymbol{\gamma}_a)]$ scales polynomially with the number of modes.

Notice that, in general, this estimation algorithm is not as optimal as the estimation algorithm in terms of the $(\mathbf{s})$ -PQDs of the data-encoding states, see Eq. (10). Indeed, the negative volume at the output of a quantum circuit is a lower bound of the product of the negative volumes of circuit elements because non-classical processes may reduce the effects of each other, e.g., a squeezing process and an anti-squeezing process can cancel each other.

B Decomposition of kernel estimation

In this section, we show that for quantum-efficient encodings, i.e., $\text{Tr}[\rho_{\Pi(x)}] \geq 1/\text{poly}(m)$ and $\text{Tr}[\rho_{\Pi(x')}] \geq 1/\text{poly}(m)$, estimating the ratio $K(x, x') = \frac{\text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]}{\text{Tr}[\rho_{\Pi(x)}]\text{Tr}[\rho_{\Pi(x')}]}$ up to inverse-polynomial precision with exponentially small probability of failure can be done by estimating $\text{Tr}[\rho_{\Pi(x)}]$, $\text{Tr}[\rho_{\Pi(x')}]$ and $\text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]$ independently up to (smaller) inverse-polynomial precision with (smaller) exponentially small probability of failure, and computing the ratio of those estimates.

We rely on the following technical result:

Lemma 2. *Let $0 < \epsilon < \epsilon' < 1$, let $\delta > 0$, let $a, b, c \in [0, 1]$ and let $\tilde{a}, \tilde{b}, \tilde{c} \in \mathbb{R}$ be random variables such that*

$$|a - \tilde{a}| \leq \epsilon, \quad (43)$$

$$|b - \tilde{b}| \leq \epsilon, \quad (44)$$

$$|c - \tilde{c}| \leq \epsilon, \quad (45)$$

each with probability greater than or equal to $1 - \delta$. Finally, assume that $b > \epsilon'$ and $c > \epsilon'$. Then $b, c, \tilde{b}, \tilde{c}$ do not vanish and

$$\left| \frac{a}{bc} - \frac{\tilde{a}}{\tilde{b}\tilde{c}} \right| \leq \frac{(3 + \epsilon)\epsilon}{\epsilon'^2(\epsilon' - \epsilon)^2}, \quad (46)$$

all with probability at least $1 - 3\delta$.

Proof. With the union bound, we have

$$|a - \tilde{a}| \leq \epsilon, \quad (47)$$

$$|b - \tilde{b}| \leq \epsilon, \quad (48)$$

$$|c - \tilde{c}| \leq \epsilon, \quad (49)$$

all together with probability at least $1 - 3\delta$. In that case, $\tilde{b} \geq \epsilon' - \epsilon$ and $\tilde{c} \geq \epsilon' - \epsilon$, so $b\tilde{c}\tilde{c} \geq \epsilon'^2(\epsilon' - \epsilon)^2$. This gives

$$\left| \frac{a}{bc} - \frac{\tilde{a}}{\tilde{b}\tilde{c}} \right| = \frac{|a\tilde{b}\tilde{c} - \tilde{a}bc|}{b\tilde{c}\tilde{c}} \quad (50)$$

$$\leq \frac{|a - \tilde{a}|b\tilde{c} + a|\tilde{b} - b|\tilde{c} + a\tilde{b}|\tilde{c} - c|}{\epsilon'^2(\epsilon' - \epsilon)^2} \quad (51)$$

$$\leq \frac{(3 + \epsilon)\epsilon}{\epsilon'^2(\epsilon' - \epsilon)^2}, \quad (52)$$

where we used the triangle inequality in the second line, and $a, b, c \in [0, 1]$ and $\tilde{b} \leq 1 + \epsilon$ in the last line. $\square$

In particular, with $\epsilon = (\epsilon'/2)^4\epsilon''$ and $\delta = \delta'/3$ this implies $\left| \frac{a}{bc} - \frac{\tilde{a}}{\tilde{b}\tilde{c}} \right| \leq \epsilon''$ with probability greater than or equal to $1 - \delta'$. Using this lemma for $a = \text{Tr}[\rho_{\Pi(x)}\rho_{\Pi(x')}]$, $b = \text{Tr}[\rho_{\Pi(x)}]$ and $c = \text{Tr}[\rho_{\Pi(x')}]$, with $\epsilon', \epsilon'' = O(1/\text{poly}(m))$ and $\delta = O(1/\exp(m))$ completes the proof.

C Phase-space quasi-probability distributions of states after loss channels

Let us consider a single-mode loss channel Λ_η with transmissivity $0 \leq \eta \leq 1$. This channel reduces the amplitude of a coherent state

$$\Lambda_\eta(|\alpha\rangle\langle\alpha|) = |\sqrt{\eta}\alpha\rangle\langle\sqrt{\eta}\alpha|. \quad (53)$$

We can expand single-mode displacement operators in terms of coherent states

$$D(\xi) = e^{|\xi|^2/2} e^{-\xi^* a} e^{\xi a^\dagger} = e^{|\xi|^2/2} e^{-\xi^* a} \frac{1}{\pi} \int d^2\alpha |\alpha\rangle\langle\alpha| e^{\xi a^\dagger} = \frac{1}{\pi} e^{|\xi|^2/2} \int d^2\alpha e^{\xi\alpha^* - \alpha\xi^*} |\alpha\rangle\langle\alpha|, \quad (54)$$

where we used the Baker–Campbell–Hausdorff formula, the resolution of identity in terms of coherent states and $a|\alpha\rangle = \alpha|\alpha\rangle$. By using this relation, Eq. (53) as well as the linearity of quantum channels, we can then find the action of loss channels on single-mode displacement operators

$$\begin{aligned} \Lambda_\eta(D(\xi)) &= \frac{1}{\pi} e^{|\xi|^2/2} \int d^2\alpha e^{\xi\alpha^* - \alpha\xi^*} \Lambda_\eta(|\alpha\rangle\langle\alpha|) \\ &= \frac{1}{\pi} e^{|\xi|^2/2} \int d^2\alpha e^{\xi\alpha^* - \alpha\xi^*} |\sqrt{\eta}\alpha\rangle\langle\sqrt{\eta}\alpha| \\ &= \frac{1}{\pi} e^{|\xi|^2/2} \int \frac{d^2\beta}{\eta} e^{\xi\beta^*/\sqrt{\eta} - \xi^*\beta/\sqrt{\eta}} |\beta\rangle\langle\beta| \\ &= \frac{1}{\eta} e^{|\xi|^2/2} e^{-\xi^* a/\sqrt{\eta}} \frac{1}{\pi} \int d^2\beta |\beta\rangle\langle\beta| e^{\xi a^\dagger/\sqrt{\eta}} \\ &= \frac{1}{\eta} \exp\left[-\left(\frac{1}{\eta} - 1\right) \frac{|\xi|^2}{2}\right] D(\xi/\sqrt{\eta}). \end{aligned} \quad (55)$$

The adjoint map Λ_η^* is related to Λ_η through this relation

$$\text{Tr}[\Lambda_\eta(D(\xi))D(\zeta)] = \text{Tr}[D(\xi)\Lambda_\eta^*(D(\zeta))]. \quad (56)$$

Thus, using $\text{Tr}[D(\xi)D(\eta)] = \pi\delta^2(\xi + \eta)$, we have

$$\begin{aligned} \text{Tr}[\Lambda_\eta(D(\xi))D(\zeta)] &= \frac{1}{\eta} \exp\left[-\left(\frac{1}{\eta} - 1\right) \frac{|\xi|^2}{2}\right] \text{Tr}[D(\xi/\sqrt{\eta})D(\zeta)] \\ &= \frac{1}{\eta} \exp\left[-\left(\frac{1}{\eta} - 1\right) \frac{|\xi|^2}{2}\right] \pi\delta^2(\xi/\sqrt{\eta} + \zeta) \\ &= \exp\left[-\left(\frac{1}{\eta} - 1\right) \frac{|\xi|^2}{2}\right] \pi\delta^2(\xi + \zeta\sqrt{\eta}) \\ &= \exp\left[-(1 - \eta) \frac{|\zeta|^2}{2}\right] \pi\delta^2(\xi + \zeta\sqrt{\eta}) \\ &= \exp\left[-(1 - \eta) \frac{|\zeta|^2}{2}\right] \text{Tr}[D(\xi)D(\zeta\sqrt{\eta})] \\ &= \text{Tr}[D(\xi)\Lambda_\eta^*(D(\zeta))]. \end{aligned} \quad (57)$$

This implies that the action of the adjoint map on displacement operators is given by

$$\Lambda_\eta^*(D(\zeta)) = \exp\left[-(1 - \eta) \frac{|\zeta|^2}{2}\right] D(\zeta\sqrt{\eta}). \quad (58)$$

This in turn gives us the action of the adjoint map on the frame operators (9) defining the single-mode (s)-PQDs,

$$\Lambda_\eta^*(\Delta^{(s)}(\alpha)) = \int_{\mathbb{C}} \frac{d^2\xi}{\pi^2} e^{-(1-\eta)|\xi|^2/2} D(\sqrt{\eta}\xi) e^{s|\xi|^\dagger/2} e^{\alpha\xi^* - \xi\alpha^*} = \frac{1}{\eta} \Delta^{(s/\eta - (1-\eta)/\eta)}(\alpha/\sqrt{\eta}). \quad (59)$$

Employing this relation, we can express the (s)-PQD of the quantum state after loss $\rho = \Lambda_\eta(\rho_{\text{in}})$ as

$$W_\rho^{(s)}(\alpha) = \text{Tr}[\Lambda_\eta(\rho_{\text{in}})\Delta^{(s)}(\alpha)] = \text{Tr}[\rho_{\text{in}}\Lambda_\eta^*(\Delta^{(s)}(\alpha))] = \frac{1}{\eta} W_{\rho_{\text{in}}}^{(s/\eta - (1-\eta)/\eta)}(\alpha/\sqrt{\eta}), \quad (60)$$

where $W_{\rho_{\text{in}}}^{(s)}(\alpha)$ is the (s) -PQD of the initial state ρ_{in} .

Considering the tensor product of single-mode loss channels $\Lambda_{\boldsymbol{\eta}} = \Lambda_{\eta_1} \otimes \cdots \otimes \Lambda_{\eta_m}$, Eq. (59) can be generalized for multimode $\Delta^{(s)}(\alpha)$, defined by Eq. (9),

$$\Lambda_{\boldsymbol{\eta}}^*(\Delta^{(s)}(\alpha)) = \int_{\mathbb{C}^m} \frac{d^{2m}\boldsymbol{\xi}}{\pi^{2m}} e^{-\boldsymbol{\xi}(I-\boldsymbol{\eta})\boldsymbol{\xi}^\dagger/2} D(\boldsymbol{\xi}\sqrt{\boldsymbol{\eta}}) e^{\boldsymbol{\xi}s\boldsymbol{\xi}^\dagger/2} e^{\alpha\boldsymbol{\xi}^\dagger-\boldsymbol{\xi}\alpha^\dagger} = \frac{1}{\det \boldsymbol{\eta}} \Delta^{(\boldsymbol{\eta}^{-1/2}(s-I+\boldsymbol{\eta})\boldsymbol{\eta}^{-1/2})}(\alpha\boldsymbol{\eta}^{-1/2}), \quad (61)$$

where $\boldsymbol{\eta} = \text{diag}(\eta_1, \dots, \eta_m)$. By using this relation, the (s) -PQD of the state after an m -mode loss channel $\rho = \Lambda_{\boldsymbol{\eta}}(\rho_{\text{in}})$ can be expressed in terms of the (t) -PQD of the m -mode initial state ρ_{in}

$$W_{\rho}^{(s)}(\alpha) = \frac{1}{\det \boldsymbol{\eta}} W_{\rho_{\text{in}}}^{(t)}(\alpha\boldsymbol{\eta}^{-1/2}), \quad (62)$$

where $\boldsymbol{t} = \boldsymbol{\eta}^{-1/2}(s-I+\boldsymbol{\eta})\boldsymbol{\eta}^{-1/2}$, or equivalently $\boldsymbol{s} = \boldsymbol{\eta}^{1/2}\boldsymbol{t}\boldsymbol{\eta}^{1/2} + I - \boldsymbol{\eta}$. These reduce to the expression given in the main text in the case of diagonal matrices of ordering parameters.

D Estimation of lossy photonic quantum kernels using Gurvits algorithm

In this section, we derive an alternative approach to classical estimation of quantum kernels based on lossy single-photon states fed into LONs based on Gurvits's algorithm for estimating the permanent [34].

In this case, the kernel function takes the form

$$K(x, x') = \text{Tr} \left[U(x) \bigotimes_{j=1}^m ((1-\eta_j)|0\rangle\langle 0| + \eta_j|1\rangle\langle 1|) U(x)^\dagger U(x') \bigotimes_{j=1}^m ((1-\eta_j)|0\rangle\langle 0| + \eta_j|1\rangle\langle 1|) U(x')^\dagger \right] \quad (63)$$

$$= \sum_{\boldsymbol{p}, \boldsymbol{q} \in \{0,1\}^m} \prod_{j=1}^m f_{\eta_j}(p_j) f_{\eta_j}(q_j) |\langle p_1 \dots p_m | V(x, x') | q_1 \dots q_m \rangle|^2, \quad (64)$$

where we have defined $V(x, x') := U(x)^\dagger U(x')$ and $f_{\eta}(p) := \eta^{1-p}(1-\eta)^p$. This means that the kernel function is equal to the expectation value of $|\langle \boldsymbol{p} | V(x, x') | \boldsymbol{q} \rangle|^2$ for $p_1, \dots, p_m$ and $q_1, \dots, q_m$ both drawn from the product of univariate Bernoulli distributions over $\{0, 1\}$ with probability η_j . Combined with Gurvits's algorithm for estimating the permanent [34], this readily gives a classical estimation algorithm for the kernel function:

- For $j \in \{1, \dots, m\}$, sample a bit p_j from the univariate Bernoulli distributions over $\{0, 1\}$ with probability η_j .
- For $j \in \{1, \dots, m\}$, sample a bit q_j from the univariate Bernoulli distributions over $\{0, 1\}$ with probability η_j .
- If $\|\boldsymbol{p}\|_1 \neq \|\boldsymbol{q}\|_1$ output 0 and halt. Otherwise, let $n = \|\boldsymbol{p}\|_1 = \|\boldsymbol{q}\|_1$.
- Let $V_n(x, x')$ be the $n \times n$ matrix obtained from $V(x, x')$ by deleting the j^{th} row (resp. j^{th} column) if $p_j = 0$ (resp. $q_j = 0$) for all $j \in \{1, \dots, m\}$.
- Let $W(x, x') := V_n(x, x') \oplus V_n(x, x')^*$, such that $\text{Per}[W(x, x')] = |\text{Per}[V_n(x, x')]|^2$. We write $W(x, x') = (w_{ij}(x, x'))_{1 \leq i, j \leq 2n}$.
- Sample uniformly N bit-strings $(y_1^{(l)}, \dots, y_{2n}^{(l)}) \in \{-1, 1\}^{2n}$ for $l \in \{1, \dots, N\}$.
- Output $\frac{1}{N} \sum_{l=1}^N y_1^{(l)} \dots y_{2n}^{(l)} \prod_{i=1}^{2n} \sum_{j=1}^{2n} y_j^{(l)} w_{ij}(x, x')$.

By Hoeffding's inequality [33] and given that the above estimator is bounded by $\|W(x, x')\|^{2n} \leq 1$ since $W(x, x')$ is the submatrix of a unitary matrix, the estimate obtained is an ϵ -close additive estimate of the kernel function with probability at least $1 - \delta$ whenever $N \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2}{\delta}\right)$.

In case all η_j 's are equal to some $\eta \in (0, 1)$, the first three steps of the above procedure can be replaced by the following ones:

- Compute $\theta = \sum_{n=0}^m \binom{m}{n}^2 \eta^{2n} (1 - \eta)^{2(m-n)} \in [0, 1]$ and sample $b \in \{0, 1\}$ from a Bernoulli distribution with parameter θ , i.e., the probability of $b = 1$ equals θ .
- If $b = 0$ output 0 and halt. Otherwise, if $b = 1$, sample $n \in \{0, 1, \dots, m\}$ from the binomial distribution with parameters (m, η) , i.e., the probability of picking n equals $\binom{m}{n} \eta^n (1 - \eta)^{m-n}$.
- Sample vectors $\mathbf{p}, \mathbf{q} \in \{0, 1\}^m$ independently and uniformly at random under the constraint $\|\mathbf{p}\|_1 = \|\mathbf{q}\|_1 = n$ ¹.

E Classical estimation of quantum kernel functions for partially measured Gaussian states

In this section, we give a proof of Theorem 1, which we recall below:

Theorem 1. *For any classical data x , let $\rho(x)$ be a quantum state encoding over m modes obtained by performing a possibly non-Gaussian measurement of the first k modes of a $(k + m)$ -mode Gaussian state $\rho_G(x)$, as in Eq. (36). Let $\tau(x)$ denote the non-classical depth of $\rho_G(x)$ (see Eq. (35) and [23]) and let $\tau(x, x') = \max(\tau(x), \tau(x')) \in [0, \frac{1}{2}]$. Then, assuming that the encoding is quantum-efficient, Algorithm 2 provides an estimate of the quantum kernel $K(x, x') = \text{Tr}[\rho(x)\rho(x')]$ with additive precision ϵ and success probability $1 - \delta$ in time*

$$O\left(\frac{\log(\frac{2}{\delta})\text{poly}(m)}{\epsilon^2(1 - \tau(x, x'))^{4k+2}}\right). \quad (65)$$

In particular, this provides an efficient classical algorithm for quantum kernel estimation whenever $k = O(\log m)$ or $\tau(x, x') = O(\log m/k)$.

Proof. Following Algorithm 2, given two Gaussian states $\rho_G(x)$ and $\rho_G(x')$ over $k + m$ modes we define the state $\sigma(x, x')$ as the $(2k)$ -mode Gaussian state obtained by taking the partial overlap of the last m modes of $\rho_G(x)$ and $\rho_G(x')$ (see Fig. 3). We also denote by $\sigma(x)$ and $\sigma(x')$ the reduced states $\text{Tr}_{k+1\dots k+m}[\rho_G(x)]$ and $\text{Tr}_{k+1\dots k+m}[\rho_G(x')]$, respectively.

Algorithm 2 combines three independent subroutines that have the same structure as Algorithm 1 (see Figs. 2 and 3). We have shown in Appendix B that if each subroutine is efficient, then Algorithm 2 is also efficient for quantum-efficient encoding.

We now bound the complexity of each subroutine using Eqs. (20) and (21). We obtain that the total number of samples for classical estimation up to additive precision ϵ and success probability $1 - \delta$ is given by

$$N \geq \frac{2}{\epsilon^2} \left[\mathcal{R}(E(x))^2 + \mathcal{R}(E(x'))^2 + \mathcal{R}(E(x, x'))^2 \right] \ln\left(\frac{2}{\delta}\right), \quad (66)$$

where

$$\mathcal{R}(E(x)) = \mathcal{N}(W_{\sigma(x)}^{(s)}) \mathcal{R}(W_{\Pi(x)}^{(-s)}) \quad (67)$$

$$\mathcal{R}(E(x')) = \mathcal{N}(W_{\sigma(x')}^{(s')}) \mathcal{R}(W_{\Pi(x')}^{(-s')}) \quad (68)$$

$$\mathcal{R}(E(x, x')) = \mathcal{N}(W_{\sigma(x, x')}^{(\mathbf{u} \oplus -\mathbf{v})}) \mathcal{R}(W_{\Pi(x) \otimes \Pi(x')}^{(-\mathbf{u} \oplus \mathbf{v})}). \quad (69)$$

¹Here is a method for sampling a uniform $\mathbf{p}$ satisfying $\|\mathbf{p}\|_1 = n$. First, let $p_1 = 1$ with probability n/m . If $p_1 = 1$, then let $p_2 = 1$ with probability $(n - 1)/m - 1$; if $p_1 = 0$, then let $p_2 = 1$ with probability $n/(m - 1)$. Continue recursively with the rest of coordinates.

We choose the largest possible ordering parameters of the form $\mathbf{s} = s\mathbf{I}$ for each of the $(\mathbf{s})$ -PQDs of the states $\sigma(x)$, $\sigma(x')$ and $\sigma(x, x')$ in all three subroutines of the algorithm, such that the corresponding $(\mathbf{s})$ -PQDs are non-negative. Writing these parameters s , s' and t , respectively, this gives

$$\mathcal{R}(E(x)) = \mathcal{R}(W_{\Pi(x)}^{(-s\mathbf{I})}) \quad (70)$$

$$\mathcal{R}(E(x')) = \mathcal{R}(W_{\Pi(x')}^{(-s'\mathbf{I})}) \quad (71)$$

$$\mathcal{R}(E(x, x')) = \mathcal{R}(W_{\Pi(x) \otimes \Pi(x')}^{(-t\mathbf{I})}), \quad (72)$$

where by Eq. (24),

$$\mathcal{R}(W_{\Pi(x)}^{(-s\mathbf{I})}) \leq \left(\frac{2}{s+1} \right)^{k+1} \quad (73)$$

$$\mathcal{R}(W_{\Pi(x')}^{(-s'\mathbf{I})}) \leq \left(\frac{2}{s'+1} \right)^{k+1} \quad (74)$$

$$\mathcal{R}(W_{\Pi(x) \otimes \Pi(x')}^{(-t\mathbf{I})}) \leq \left(\frac{2}{t+1} \right)^{2k+1}. \quad (75)$$

To conclude the proof, we simply need to show that the ordering parameters s, s', t may be all chosen arbitrarily close to $1 - 2\tau(x, x')$, where $\tau(x, x')$ is the maximal non-classical depth of $\rho_G(x)$ and $\rho_G(x')$. To do so, we prove the following properties of the non-classical depth of Gaussian states, which appear to be new:

Lemma 3. *The non-classical depth of Gaussian states is non-increasing under partial trace and non-increasing under partial overlap.*

Proof. Let σ be a Gaussian state over $k+m$ modes with covariance matrix Σ and displacement vector $\bar{\mathbf{r}}$. Recall the expression for the $(\mathbf{s})$ -PQD of a Gaussian state from Eq. (35):

$$W_{\sigma}^{(s)}(\boldsymbol{\alpha}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\alpha} - \bar{\mathbf{r}})(\Sigma - \mathbf{s} \oplus \mathbf{s})^{-1}(\boldsymbol{\alpha} - \bar{\mathbf{r}})^{\top}}}{(2\pi)^{m+k} \sqrt{\det(\Sigma - \mathbf{s} \oplus \mathbf{s})}}, \quad (76)$$

for all $\boldsymbol{\alpha} \in \mathbb{C}^{k+m}$ and all $\mathbf{s}$ such that $\Sigma - \mathbf{s} \oplus \mathbf{s}$ is positive definite. By Definition 1, the non-classical depth of the Gaussian state σ is thus given by $\tau = \frac{1}{2}(1 - s)$, where s is the supremum of the values such that $\Sigma \succ s\mathbf{I}$, with $\mathbf{I}$ being the $(2k+2m) \times (2k+2m)$ identity operator.

Let us write

$$\Sigma = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}. \quad (77)$$

where A is a $(2k) \times (2k)$ symmetric matrix and C is a $(2m) \times (2m)$ symmetric matrix. The condition $\Sigma \succ s\mathbf{I}$ is equivalent to

$$X^T A X + Y^T C Y + 2X^T B^T Y > s(X^T X + Y^T Y), \quad (78)$$

for all $(X, Y) \in \mathbb{R}^{2k} \times \mathbb{R}^{2m}$. Setting $X = 0$ gives $Y^T C Y > sY^T Y$ for all $Y \in \mathbb{R}^{2m}$, and thus $C \succ s\mathbf{I}$, which implies that the non-classical depth of the Gaussian state obtained by taking the partial trace of σ over the first k modes (with covariance matrix C [56]) is smaller than that of σ . This shows that the non-classical depth of a Gaussian state is non-increasing under partial trace.

Let us now consider an additional Gaussian state σ' , with covariance matrix Σ' . The partial overlap over the last m modes of σ and σ' is defined as

$$\sigma'' := \text{Tr}_{k+1, \dots, k+m, 2k+m+1, \dots, 2k+2m} \left[(\sigma^T \otimes \sigma') \bigotimes_{j=1}^m |\text{TWB}\rangle \langle \text{TWB}|_{k+j, 2k+m+j} \right], \quad (79)$$

where $|\text{TWB}\rangle := \sum_{n \geq 0} |nn\rangle$ is a infinitely-squeezed two-mode squeezed state or twin-beam state (TWB). The operator $|\text{TWB}\rangle\langle\text{TWB}|$ can equivalently be obtained by sending a position eigenstate and a momentum eigenstate into a balanced beam-splitter [37], so the partial overlap can be expressed as

$$\sigma'' = \text{Tr}_{k+1, \dots, k+m, 2k+m+1, \dots, 2k+2m} [U_{BS}(\sigma^T \otimes \sigma') U_{BS}^\dagger \times (\mathbb{I}_k \otimes |0\rangle\langle 0|_{q_{k+1}} \otimes \dots \otimes |0\rangle\langle 0|_{q_{k+m}} \otimes \mathbb{I}_k \otimes |0\rangle\langle 0|_{p_{2k+m+1}} \otimes \dots \otimes |0\rangle\langle 0|_{p_{2k+2m}})], \quad (80)$$

where $U_{BS} = \bigotimes_{j=1}^m U_{k+j, 2k+m+j}$ is the passive linear operator corresponding to the action of the balanced beam splitters, and where q_j and p_j denote the position and momentum quadrature operators for the j^{th} mode, respectively.

For any symmetric matrices M and M' , $M \succ sI$ implies $OMO^T \succ sI$ for any orthogonal matrix O , and $M \succ sI$ and $M' \succ s'I$ implies $M \oplus M' \succ \min(s, s')I$. The covariance matrix of the Gaussian state $\sigma''' := U_{BS}(\sigma^T \otimes \sigma') U_{BS}^\dagger$ is given by $\Sigma''' := S_{U_{BS}}(T\Sigma T \oplus \Sigma') S_{U_{BS}}^T$, where $S_{U_{BS}}$ is the orthogonal matrix corresponding to the action of U_{BS} on the vector of quadrature operators and $T = I \oplus (-I)$ is the orthogonal matrix corresponding to the action of the transposition on the vector of quadrature operators. In particular, its classical depth is smaller than the maximum of the non-classical depths of σ and σ' .

Finally, we write the covariance matrix of $\sigma''' = U_{BS}(\sigma^T \otimes \sigma') U_{BS}^\dagger$ as

$$\Sigma''' = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}, \quad (81)$$

where A is a $(4k) \times (4k)$ symmetric matrix and C is a $(4m) \times (4m)$ symmetric matrix, and where we have ordered the vector of quadrature operators as (to get more convenient expressions later on):

$$\mathbf{r} = (q_1, \dots, q_k, p_1, \dots, p_k, q_{k+m+1}, \dots, q_{2k+m}, p_{k+m+1}, \dots, p_{2k+m}, q_{k+1}, \dots, q_{k+m}, p_{2k+m+1}, \dots, p_{2k+2m}, p_{k+1}, \dots, p_{k+m}, q_{2k+m+1}, \dots, q_{2k+2m})^T. \quad (82)$$

Then, from Eq. (80) the covariance matrix of the partial overlap state σ'' is the conditional covariance matrix corresponding to a measurement of the position quadratures for the modes $k+1, \dots, k+m$ and of the momentum quadratures for the modes $2k+m+1, \dots, 2k+2m$, that is [56]

$$\Sigma'' = A - B(\Pi C \Pi)^{-1} B^T, \quad (83)$$

where $\Pi = I_{2m} \oplus 0_{2m}$ is the projector selecting the quadratures being measured, and where the inverse is understood in the generalized (pseudo-inverse) sense, i.e., $(\Pi C \Pi)^{-1} = C_1^{-1} \Pi$, where C_1 is the top-left block of C selected by Π . Then, the condition $\Sigma'' \succ sI$ implies $C - sI \succ 0$, so $C - sI$ is invertible, and

$$\begin{pmatrix} A - sI & B \\ B^T & C - sI \end{pmatrix} \succ 0. \quad (84)$$

Hence, the Schur complement of $C - sI$ in this matrix is also positive definite [57], i.e.,

$$A - sI - B(C - sI)^{-1} B^T \succ 0. \quad (85)$$

Now for all $s \geq 0$ we have

$$(C - sI)^{-1} \succeq C^{-1}. \quad (86)$$

Writing $\begin{pmatrix} C_1 & C_2 \\ C_2^T & C_3 \end{pmatrix}$, for all $(X, Y) \in \mathbb{R}^{2m} \times \mathbb{R}^{2m}$ we have

$$(X^T \ Y^T) [C^{-1} - (\Pi C \Pi)^{-1}] \begin{pmatrix} X \\ Y \end{pmatrix} = X^T C_1^{-1} C_2 S^{-1} C_2^T C_1^{-1} X - 2Y^T S^{-1} C_2^T C_1^{-1} X + Y^T S^{-1} Y, \quad (87)$$

where $S = C_3 - C_2^T C_1^{-1} C_2 \succ 0$ is the Schur complement of $C_1 \succ 0$, and where we have used

$$(\Pi C \Pi)^{-1} = \begin{pmatrix} C_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad (88)$$

and the block inversion formula [58]

$$C^{-1} = \begin{pmatrix} C_1 & C_2 \\ C_2^T & C_3 \end{pmatrix}^{-1} = \begin{pmatrix} C_1^{-1} + C_1^{-1} C_2 S^{-1} C_2^T C_1^{-1} & -C_1^{-1} C_2 S^{-1} \\ -S^{-1} C_2^T C_1^{-1} & S^{-1} \end{pmatrix}. \quad (89)$$

Setting $X' := C_2^T C_1^{-1} X$ in Eq. (87) we obtain

$$\begin{pmatrix} X^T & Y^T \end{pmatrix} [C^{-1} - (\Pi C \Pi)^{-1}] \begin{pmatrix} X \\ Y \end{pmatrix} = X'^T S^{-1} X' - 2Y^T S^{-1} X' + Y^T S^{-1} Y \quad (90)$$

$$= (X' - Y)^T S^{-1} (X' - Y) \geq 0, \quad (91)$$

since $S^{-1} \succ 0$. This implies $C^{-1} \succeq (\Pi C \Pi)^{-1}$ and together with Eq. (86) we obtain $(C - sI)^{-1} \succeq (\Pi C \Pi)^{-1}$ and thus $B(C - sI)^{-1} B^T \succeq B(\Pi C \Pi)^{-1} B^T$. With Eq. (85), this finally yields

$$\Sigma'' = A - B(\Pi C \Pi)^{-1} B^T \succeq A - B(C - sI)^{-1} B^T \succ sI, \quad (92)$$

when assuming that $\Sigma''' \succ sI$. This shows that the non-classical depth of σ'' is smaller than that of σ''' , which itself was smaller than the maximum of the non-classical depths of σ and σ' . This completes the proof that the non-classical depth of Gaussian states is non-increasing under partial overlap. $\square$

Since $\sigma(x)$ (resp. $\sigma(x')$) is obtained from $\rho_G(x)$ (resp. $\rho_G(x')$) by taking a partial trace, and $\sigma(x, x')$ is a partial overlap of the states $\rho_G(x)$ and $\rho_G(x')$, Lemma 3 ensures that the non-classical depths of the states $\sigma(x), \sigma(x'), \sigma(x, x')$ are all bounded by $\tau(x, x')$, the maximal non-classical depth of $\rho_G(x)$ and $\rho_G(x')$. By Definition 1, this implies that the ordering parameters s, s', t in Eq. (73) may be all chosen arbitrarily close to $1 - 2\tau(x, x')$, which concludes the proof of Theorem 1, by noting that all the covariance matrices involved can be computed in time $\text{poly}(m)$. $\square$

F Kernel estimation for adaptive Gaussian boson sampling

Given the limitations of quantum computations based on LONs for quantum kernel methods identified in the main text, we can ask whether simple extensions of LONs can restore their usefulness. One such natural extension is through the addition of *adaptivity* in the measurements. Also known as feed-forward, adaptivity refers to the possibility of modifying part of the computation based on the outcomes of intermediate measurements, as in Measurement-Based Quantum Computing [59] in the quantum circuit picture. Adaptivity is particularly relevant in the context of quantum computing with LONs, as it allows for performing universal quantum computations, e.g. through the Knill–Laflamme–Milburn scheme [60] or the more recent Fusion-Based Quantum Computing model [61]. In those schemes, the addition of adaptive measurements to LONs allows for the active switching of offline resource entangled states into a LON, which can be used to implement a universal gate set on qubits encoded using photons in a near-deterministic fashion.

Kernel estimation becomes BQP-complete in the regime of enough adaptive measurements $k = \text{poly}(m)$, where $m = \text{poly}(n)$ is the number of photonic modes supporting the computations over n qubits. Hence, it is expected that, unless $\text{BPP} = \text{BQP}$, estimating quantum kernels that are based on LONs with adaptive measurements is hard for classical computers (note that hardness of kernel estimation does not necessarily entails hardness of the corresponding learning task).

What are the conditions necessary to enable quantum computational advantage through quantum kernel methods with adaptive measurements? This question was considered in [62] for adaptive boson

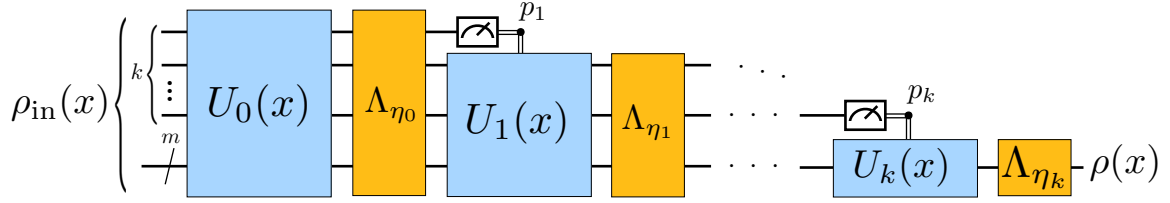


Figure 4: An adaptive Gaussian boson sampling computation, parametrised by classical data x . A total of k modes of the initial $k + m$ modes are being adaptively measured with photon-number detection. Specific photon number measurement outcomes are being represented here, but our results cover general adaptive measurement strategies. The unitary operators $U_0(x), \dots, U_k(x)$ are Gaussian. The yellow blocks $\Lambda_{\eta_0}, \dots, \Lambda_{\eta_k}$ represent potential layers of losses which may be non-uniform over the modes.

sampling with input Fock states, where it was shown that if too few photons are being detected by the adaptive measurements, then there exists an efficient classical algorithm for estimating quantum kernel functions. However, since near-indistinguishable single-photon states are hard to generate experimentally in a near-deterministic fashion, a natural question is to investigate the limitations of near-term quantum computational advantages through quantum kernel methods using adaptive LONs with realistic input quantum states. In particular, we consider the case of adaptive LONs with Gaussian input states (see Fig. 4).

This mirrors the evolution of quantum computational advantage experiments based on sampling from the output distribution of LONs, which have progressively shifted from proof-of-concept demonstration of boson sampling with input Fock states [9, 63] to Gaussian boson sampling [43, 64], where the input Fock states are replaced by Gaussian states, much easier to generate experimentally in optical platforms. Since single-photon states can be prepared in an offline fashion using Gaussian two-mode squeezed states and heralded photon-number measurements, and given that kernel estimation with linear optical computations using input single photons and adaptive measurements is BQP-complete, kernel estimation with linear optical computations using adaptive Gaussian boson sampling is also BQP-complete.

In what follows, we derive sufficient conditions for efficient classical estimation of quantum kernel functions based on adaptive Gaussian boson sampling output states. In particular, we show that, similar to the case of adaptive boson sampling [62], if too few photons are being detected by the adaptive measurements, then there exists an efficient classical algorithm for estimating quantum kernel functions based on adaptive Gaussian boson sampling.

By deferring the photon-number measurements to the end, the m -mode output state of a generic adaptive Gaussian boson sampling computation with k adaptive measurements takes the form

$$\rho = \sum_{\mathbf{p}} \text{Tr}_k \left[(\Pi_{\mathbf{p}} \otimes \mathbb{I}_m) U^{(\mathbf{p})} \rho_{\text{in}} U^{(\mathbf{p})\dagger} \right], \quad (93)$$

where the sum is over adaptive photon-number measurement patterns $\mathbf{p} = (p_1, \dots, p_k)$, with $\Pi_{\mathbf{p}} = |\mathbf{p}\rangle\langle\mathbf{p}|$, where the partial trace is over the first k modes, and where

$$U^{(\mathbf{p})} := \left(\mathbb{I}_k \otimes U_k^{(p_k)} \right) \left(\mathbb{I}_{k-1} \otimes U_{k-1}^{(p_{k-1})} \right) \cdots \left(\mathbb{I}_1 \otimes U_1^{(p_1)} \right) U_0, \quad (94)$$

where each U_j is a Gaussian unitary over $m + k - j$ modes depending on the previous adaptive measurement outcome p_j (see Fig. 4). We restrict to Gaussian unitary operators, but a similar reasoning extends straightforwardly to Gaussian channels.

Parametrizing the input Gaussian state ρ_{in} and the intermediate Gaussian unitary operations with classical data x, x' , the corresponding quantum kernel functions take the form

$$K(x, x') = \text{Tr}[\rho(x)\rho(x')] = \sum_{\mathbf{p}, \mathbf{p}'} \text{Tr}[\rho_{\mathbf{p}}(x)\rho_{\mathbf{p}'}(x')], \quad (95)$$

where

$$\rho_{\mathbf{p}}(x) := \text{Tr}_k \left[(\Pi_{\mathbf{p}} \otimes \mathbb{I}_m) U^{(\mathbf{p})}(x) \rho_{\text{in}}(x) U^{(\mathbf{p})}(x)^\dagger \right], \quad (96)$$

is a (sub-normalised) post-measurement state. The kernel thus rewrites as

$$K(x, x') = \sum_{\mathbf{p}, \mathbf{p}'} K_{\mathbf{p}, \mathbf{p}'}(x, x'), \quad (97)$$

with $K_{\mathbf{p}, \mathbf{p}'}(x, x') := \text{Tr}[\rho_{\mathbf{p}}(x) \rho_{\mathbf{p}'}(x')]$. Any such sub-kernel $K_{\mathbf{p}, \mathbf{p}'}(x, x')$ can be efficiently estimated through the first step of Algorithm 2 under the same conditions as in Theorem 1, i.e., if the number of adaptive measurements or the non-classical depth of the Gaussian states involved is small enough. This provides in turn a simple classical algorithm for estimating the full quantum kernel $K(x, x')$: writing $\mathcal{S}(x, x')$ a set of likely adaptive measurement patterns $\mathbf{p}, \mathbf{p}'$, i.e., which by definition satisfies

$$\Pr[(\mathbf{p}, \mathbf{p}') \notin \mathcal{S}(x, x')] \leq \frac{1}{\text{poly}(m)}, \quad (98)$$

one may sample uniformly $(\mathbf{p}, \mathbf{p}') \in \mathcal{S}(x, x')$ and use Algorithm 2 to provide an estimate $\tilde{K}$ of $K_{\mathbf{p}, \mathbf{p}'}(x, x')$ up to additive precision ϵ with failure probability δ . Then, by construction,

$$|K(x, x') - |\mathcal{S}(x, x')| \tilde{K}| \leq \epsilon + \frac{1}{\text{poly}(m)}. \quad (99)$$

with probability $1 - \delta$.

In particular, when the number of photons being detected by the adaptive measurements is too small, i.e., when $|\mathcal{S}(x, x')| \leq \text{poly}(m)$, this provides an efficient classical estimation algorithm for $K(x, x')$ under the same conditions as in Theorem 1, i.e., if the number of adaptive measurements or the non-classical depth of the Gaussian states involved is small enough.

Note that the condition $|\mathcal{S}(x, x')| \leq \text{poly}(m)$ may be checked efficiently based on the energy of the Gaussian states $U^{(\mathbf{p})}(x) \rho_{\text{in}}(x) U^{(\mathbf{p})}(x)^\dagger$ and $U^{(\mathbf{p}')}(\mathbf{x}') \rho_{\text{in}}(\mathbf{x}') U^{(\mathbf{p}')}(\mathbf{x}')^\dagger$. Moreover, $\mathcal{S}(x, x')$ may be chosen with an efficient classical description by picking the smallest $N(k, m, x), N'(k, m, x')$ such that $\mathcal{S}(N, N') := \{(\mathbf{p}, \mathbf{p}'), \text{ s.t. } |\mathbf{p}| \leq N, |\mathbf{p}'| \leq N'\}$ is a set of likely adaptive measurement patterns. A similar proof works for general POVM elements.